MSc in International Finance
Academic Year 2023-2024

# Master Research Paper

# Positive Conditional Bias in Earnings Forecasts: A Machine Learning Approach

**Arianna Morè**
**Nicolò Maria Marsucco Fiazza**

Supervisor: Prof. Thierry Foucault

5th June 2024

*PUBLIC REPORT*

*We would like to thank Professor Thierry Foucault for all the guidance and feedback received. We also extend our gratitude to our families, friends, and colleagues who have supported our efforts.*

**Abstract**

The accuracy of corporate earnings forecasts is essential for making informed financial decisions. While analysts' predictions are commonly used for these forecasts, their accuracy remains a concern. This research addresses this issue by leveraging a rolling-window random forest model from Van Binsbergen et al. (2023). This model integrates analysts' forecasts with other financial data to enhance prediction accuracy. Our research has two main goals. First, we replicate and validate the findings of the paper mentioned above. We confirm the presence of a positive conditional bias, which means analysts tend to be overly optimistic, especially for longer forecasts. Second, we explore the factors influencing the conditional bias and identify the importance of information availability and regulations. Our findings suggest that greater analyst involvement and effective regulations are associated with reduced bias. This opens doors for further research on analyst behavior, regulatory effectiveness, and the development of even more accurate forecasting methods.

**KEY WORDS**: Earnings Forecasting, Conditional Bias, Random Forest Regression, Fair Disclosure Regulation, Information Asymmetry

# Contents

# I. Introduction

Numerous finance theories suggest that the equity value of a company is intrinsically linked to the cash flows expected by its shareholders (see Koller, Goedhart, Wessels et al. (2010)). However, practitioners often use companies' earnings as a proxy for these cash flows. Therefore, one may immediately see that accurately forecasting firms' earnings is an essential input to calculate their equity value. Motivated by the importance of this real-life application, there have been numerous academic attempts to make these forecasts as accurate as possible. Nonetheless, as documented by Kothari et al. (2016) in an extensive review of these methods, their out-of-sample performance tends to be quite poor.

In an effort to improve corporate earnings predictions, Van Binsbergen et al. (2023) introduced a machine learning model that combines analysts' forecasts with other relevant financial factors to generate enhanced earnings predictions. The contribution of the authors is twofold. Firstly, they successfully provide a model which outputs superior out-of-sample results compared to commonly used linear earnings models. Secondly, they employ the results of their model to investigate the conditional bias present in analysts' forecasts. Crucially, the bias is *conditional*, indicating that the analysts' forecasts, which are employed as predictors, are compared to the results of the model and not to the actual ground truth values. The comparison and the further analysis aim at depicting the bias between the analysts and the statistically unbiased model.

The machine learning model used in the original paper is a random forest. The choice is justified by the ability of random forests in capturing nonlinear relationships and being robust to overfitting. As anticipated, the model predictors are the analysts' predictions, the financial ratios from the companies' balance sheets and other macroeconomic variables. The choice of using analysts' predictions as input implies that the model effectively attempts to improve these forecasts, rather than generating them from nothing.

In Van Binsbergen et al. (2023) the authors observe that the conditional bias of analysts is, on average, positive and increases as the forecast horizon becomes longer. This result is highly relevant as it implies that analysts tend to be overoptimistic, which could undermine the accuracy of their forecasts. The authors do not dive deep into the reasons behind the observed conditional bias and their study is limited to acknowledging its presence. Moreover, as of June 2024, the authors' source code has not been made public, implying that a replication based on the paper is necessary to examine, validate and further improve their findings.

In this dissertation, we firstly replicate the model of Van Binsbergen et al. (2023)[1]. We use our implementation to verify the reproducibility of their results and find that they can be reproduced within a small margin of error. In particular, our results confirm that the machine learning model described in the orginal paper can produce forecasts which are more accurate than analysts' ones. Moreover, the results also give credibility to the authors' finding that analysts' forecasts tend to be overoptimistic (i.e. positively biased) when compared to the statistically optimal benchmark obtained via the random forest regression, and their tendency towards overoptimism intensifies with longer forecast horizons. We further investigate the inner workings of our model by examining the feature importance of various predictors. This allows to compare our model to the original one, validating our replication and concluding that both models attribute high importance to the same set of features. In addition, by studying the time series evolution of analysts' conditional bias, we confirm the existence of the historical patterns observed by Van Binsbergen et al. (2023).

[1] Our source code is publicly available at `https://github.com/NicoloMarsucco/Man-vs-machine`

With the aim of expanding the investigation on the conditional bias begun by the original paper, we further analyze the results obtained by our model. We find evidence suggesting that analysts' conditional bias decreases as the information set regarding a company becomes larger. Moreover, a closer inspection of the time series plots suggests that the conditional bias is influenced by the introduction of new regulations, specifically the adoption of the Regulation on Fair Disclosure by the US Securities and Exchange Commission in 2000. We implement statistical tools to test these hypotheses and find quantitative evidence supporting our claims.

This dissertation is structured as follows. In Section II, we present a concise literature review of the most influential papers relevant to our research. In Section III, we describe the methodology we adopt to replicate the results of the original paper. Furthermore, in this section, we introduce the methodology used to further study the conditional bias. Section IV presents the results of the exploratory data analysis that we conducted before testing the replication of the model. Lastly, in Section V, we present and discuss the results obtained by replicating the model of the original paper, as well as from our further study of the conditional bias.

# II. Literature Review

The importance of forecasting firms' earnings, discussed in the previous section, makes it unsurprising that the investment community has long been interested in accurately forecasting companies' earnings (Johnson & Schmitt 1974). In this section, we present a brief review of some of the most popular conventional models used to forecast companies' earnings. By using the term *conventional*, we mean that these models only use information that is relatively easy to obtain, such as variables from financial statements or publicly known macroeconomic indicators. In other words, we do not consider models exploiting the so-called alternative datasets, such as the one using satellite imagery developed by Yu et al. (2023).

As outlined by Monahan (2018) in their comprehensive review of earnings forecasting methods, these methods fall into two main categories depending on the approach used. The first approach, the one most prominent in the early literature on the topic, is to treat the problem as a time-series problem of earnings alone. To put it differently, studies falling in this category attempt to use previous realisations of firms' earnings to forecast future earnings. This approach gives rise to what we call[2] *time-series models*, which we discuss further in Section A. However, more recent studies favor the *panel-data approach*. (as termed by Monahan (2018)). These models leverage both cross-sectional and time-series data for forecasting. We explore some examples of panel-data models in Section B. This review provides a foundation for understanding the challenges associated with forecasting firm earnings, which we summarize in Section C.

## A. Time-Series Models

This first formulation of the problem implicitly assumes that the future value of the earnings per share is a function of the past $N$ realisations of the earnings per share. Although this treatment is, on a conceptual level, fairly intuitive, the models used are often fairly complex. For example, Jarrett (1989) compares the use of the Holt-Winters method, the Box-Jenkins method, and autoregressions of de-seasonalised data. They find that, even though some of these methods perform relatively well on certain stocks, none of them can be considered to be *universally desirable* (Jarrett 1989). Another example is the work done by Albrecht et al. (1977); here, the authors compare forecasts made using the Box-Jenkins method to those made by analysts and they find that the former can be more accurate than the latter but only in very specific settings. Supporting the notion that the Box-Jenkins method does not always yield accurate results, Watts & Leftwich (1977) suggest that a random walk model tends to perform better than the Box-Jenkins method.

Further evidence comes from Callen et al. (1996), who compare the Brown-Rozeff model, the Griffin-Watts (Foster) model, and a neural network. Their findings indicate that simpler linear models (Brown-Rozeff and Griffin-Watts) outperform the neural network, suggesting that linear approaches may be more suitable for this specific time-series problem. The complexity of this forecasting problem is also emphasised by Bradshaw et al. (2012), who find that a random walk model often yields more accurate long-term forecasts than analysts' predictions. One interesting property of earnings time-series highlighted by Lipe & Kormendi (1994) is that they seem to exhibit a mean-reversion property. However, this observation is challenged by Ball & Watts (1972), who suggest that annual earnings appear to follow a submartingale.

[2]This nomenclature is also taken from Monahan (2018).

Overall, as noted by Monahan (2018), the random-walk time-series model appears to be the best for forecasting earnings using a time-series approach. This result highlights the poor performance of models built using this approach, suggesting that it may be suitable to approach differently this forecasting problem.

## B. Panel-Data Models

As already mentioned, the models using the panel-data approach attempt to forecast firms' earnings by using cross-sectional data (e.g. financial statements data, macroeconomic data, industry specific data etc.) from various time-steps. Due to significant advancements in computational power and limitations of time-series methods, recent models have shifted towards this approach (Green & Zhao 2022).

One of the most influential papers falling in this category is the one written by Fama & French (2000). In this study, the authors perform a nonlinear regression using variables such as dividend paid and difference between expected and realised earnings to try to forecast earnings. The resulting model is relatively complex and aims to exploit the mean reversion property of companies' profitability. According to Fama & French (2000), this model is capable of predicting changes in earnings caused by this mean reversion property; however, they also acknowledge that the model fails to capture all changes in earnings, especially when these are extreme. This model was then further developed by Hou et al. (2012), Li & Mohanram (2014) and finally by Hess & Wolf (2014). Nonetheless, all these models appear to be unable to consistently perform better than analysts.

In recent years, more computationally expensive machine learning models have been used. For instance, Zhang et al. (2004) propose a neural network to forecast companies' earnings using fundamental accounting variables and they compare it to various linear models. Their work suggests that neural networks tend to perform better than linear models for this type of forecasting. Cao & Parry (2009) further develop this neural network model and find that it can be made more accurate by training it using a genetic algorithm rather than through back propagation.

Another example of the application of complex machine learning models to this problem is presented by De Silva & Thesmar (2024). In their study, the authors select financial ratios, industry indicator variables, and information regarding the stock's price history as inputs to use to make their forecast. They then feed these inputs to various models to see which one performs best. More specifically, they compare the use of the following models: random walk model, elastic net, random forest and boosted trees. They find that, on average, analysts perform better than all these models considered. Encouragingly, their results align with our chosen methodology. They identify the random forest model as consistently achieving the highest accuracy among those considered. This finding strengthens the rationale for replicating the work of Van Binsbergen et al. (2023), who themselves employed a random forest to enhance the efficacy of earnings forecasts.

De Silva & Thesmar (2024) are not alone in highlighting the poor performance of complex models using a panel-data approach for forecasting firms' earnings. For example, Gerakos & Gramacy (2013) find that a random walk model tends to perform roughly as well as a model using a large set of predictors. Therefore, it appears there is no unique go-to panel-data model which consistently outperforms analysts' forecasts. Our investigation suggests that the inherent complexities of forecasting firm earnings contribute to this observed discrepancy between predicted and actual values. We explore these complexities in the next section.

## C. Challenges

Green & Zhao (2022) present a very lucid description of the challenges which make creating a reliable model for forecasting companies' earnings such a difficult task. We summarise below their description of these challenges, as they are useful to understand some of the choices we make in our work.

The first challenge, which may be deduced from the findings presented in the previous sections, is that, in spite of the numerous collective efforts to develop models to forecast earnings, these models are only capable of explaining a small part of the variations in earnings. In other words, earnings appear to be, to a significant degree, unpredictable. This may not come as a surprise given the number of factors which could affect earnings. It is worth noting that the information set used by many of the models presented tends to be backward-looking while, to forecast earnings, it would probably be useful to incorporate some forward-looking measures as well (such as analysts' forecasts). As a result, any new model attempting to forecast earnings needs to be able to deal with this challenge.

The second challenge relates to the fact firms' earnings tend to a have a low signal-to-noise ratio. Therefore, the amount of noise present in the time-series of earnings is so large that it makes it difficult to observe patterns and make accurate forecasts. Consequently, a model for predicting companies' earnings must be able to filter useful information from the high amount of noise.

The third and final challenge has to do with the model uncertainty. In other words, it is not unreasonable to suggest that, in light of an ever-changing economic, regulatory and accounting landscape, the model and its parameters may change in time. This final challenge adds a new layer of complexity to the problem and, together with the other above-mentioned factors, motivates the need for further research in the field of earnings forecasts.

# III. Methodology

This section outlines the methodology employed to replicate and extend the work presented by Van Binsbergen et al. (2023). In Section A we begin by presenting the datasets used. Then, in Section B, we illustrate the data engineering process defined to obtain a set of observations as similar as possible to the one proposed in the original paper. Afterwards, in Section C, we describe the model used by the original authors. Lastly, in Section D, we present the methodology which we employ to extend the work presented in the original paper. Put differently, Sections A to C deal with explaining how we replicate the results of the original paper, while Section D introduces our approach for further studying the conditional bias.

## A. Datasets

The first step towards replicating the original paper consists in obtaining the data from the same databases. These databases are summarised in Table 1; this table also indicates what type of data each database provides. The time period considered is the same one of the original paper, going from the first day of January 1985 to the last day of December 2019. All the data in these databases are about U.S. stocks and the U.S. economy; hence, all these variables (except for the financial ratios and some macroeconomic variables) are denominated in USD. However, to improve readability, we do not report the unit in the examples we present.

**Table 1:** Databases used and types of data obtained from each database.

| **Type of data** | **Database** |
|---|---|
| Realised earnings & analysts' forecasts | I/B/E/S |
| Monthly stock prices | CRSP |
| Financial ratios | I/B/E/S |
| Macroeconomic variables | Federal Reserve Bank of Philadelphia |

## B. Data Engineering

Transforming the datasets mentioned in Table 1 into a set of usable ones is a complicated process which requires a solid understanding of the type of data contained in each dataset. Hence, in Sections B.1 to B.4, we describe the type of data that can be obtained from each dataset and the transformations that we apply to each, following the original paper. Then, in Section B.5, we describe how we merge the various datasets into a single one to be fed to the machine learning model.

### B.1. Realised Earnings & Analysts' Forecast

This dataset provides the values of the realized (diluted) earnings per share (EPS) for each company, as well as the average forecasted (diluted) EPS. More details regarding the variables contained in this dataset are presented in Table 2, which also includes an example entry. This example highlights that, in certain instances, even though the forecast horizon is one

quarter, the forecast was effectively made less than a month before the company announced its earnings results. Similarly, not all forecasts for the two-quarter forecast horizon were made exactly two quarters before the company announced its earnings results, and so on for all the forecasting periods. Furthermore, it is also important to highlight that, when the forecast horizon is equal to one, two or three quarters, the realised earnings are reported in quarterly values, while when the forecast horizon is equal to one or two years, the realised earnings are reported in annualised values.

Since the values of realised earnings are rounded to the nearest cent, we follow Van Binsbergen et al. (2023) in using the cumulative adjustment factor `cfacshr` (obtained from the CSRP database) to try to decrease the rounding errors. Therefore, if the consensus estimate is calculated at time $t-\tau$ and the earnings are announced at time $t$, then the value of realised earnings can be adjusted as follows:

$$\texttt{actual\_adjusted}_t = \texttt{actual}_t \times \frac{\texttt{cfacshr}_{t\ \tau}}{\texttt{cfacshr}_t} \tag{1}$$

Furthermore, for reasons explained in the Exploratory Data Analysis section (Section IV), we trim the dataset by removing 1% outliers. Finally, the model in the original paper also uses the most recent earnings realisation as an input, so we create a new column (labelled `adj_past_eps`) where we insert the most recent value of realised earning for each company. Care was taken to prevent possible data leakage; we specifically made sure that the announcement date of the previous earnings was always before or equal to the date of when the average analysts' forecast was computed.

**Table 2:** Explanation of the meaning of the variables obtained from the I/B/E/S database, created using the information available at Dai (2020). The last column of this table shows a sample entry. The relationship between the forecast horizon identifier and the forecast horizon is the following: the forecast horizon is one quarter when `fpi`=6, two quarters when `fpi`=7, three quarters when `fpi`=8, one year when `fpi`=1 and two years when `fpi`=2.

| **Variable** | **Meaning** | **Example entry** |
|---|---|---|
| `ticker` | Ticker of the company | 0000 |
| `cusip` | 9-digit company identifier | 87482X10 |
| `cname` | Name of the company | Talmer Bancorp |
| `fpedats` | Fiscal period end date | 31/03/2014 |
| `statpers` | Date of when the analysts' average forecast was calculated | 17/04/2014 |
| `meanest` | Average analyst forecast of the companies' earnings | 0.08 |
| `actual` | Realised earnings (expressed as diluted EPS) | 0.12 |
| `anndats_act` | Date of when the company announces its earnings | 06/05/2014 |
| `fpi` | Forecast horizon identifier | 6 |
| `numest` | Number of analyst estimates used to compute `meanest` | 4 |

### B.2. Monthly Stock Prices

We present the meaning of the variables obtained from the CRSP database in Table 3. Just as before, this table contains an example entry in the last column.

**Table 3:** Explanation of the meaning of the variables obtained from the CRSP database and a sample entry. This table was created using the information provided by the Wharton Research Data Services.

| Variable | Meaning | Example entry |
|---|---|---|
| permno | CRSP security identifier | 10001 |
| cusip | 9-digit company identifier | 36720410 |
| date | Date of when the price was recorded (last business day of the month) | 30/10/1987 |
| ret | Most recent monthly return of the stock | 0.0200 |
| prc | Share price | 6.3750 |
| cfacshr | Cumulative adjustment factor | 3 |

### B.3. Financial Ratios

The financial ratios were also obtained from the I/B/E/S database. Because there are about 80 financial ratios for each company, we do not list the meaning of each one. Readers who wish to understand their meanings are invited to refer to the I/B/E/S documentation. Rather, we only present an example of an entry of this database to better illustrate the modifications made. This example is shown in Table 4.

The first change that we make to this dataset is to drop the financial ratios labelled PEG_1yrforward, PEG_ltgforward, pe_op_basic and pe_op_dil because they are missing too many data points. Then, we fill the missing financial ratios using the median of the same column of the companies from the the same industry group (indicated by the variable ffi49) from the same time step. For example, consider the observation shown in Table 4. Assume that this entry is missing the value for the financial ratio labelled capital_ratio. Then, to fill this value, we would use the median of the values of capital_ratio of all the tech companies (ffi=35) which reported their earnings in September 2000 (same month as public_date).

Finally, for each company, we interpolate in time the values of the financial ratios. To do this, we assume that all the financial ratios remain constant until the company has a new earning announcement.

**Table 4:** Example of an entry obtained from the dataset containing the financial ratios from the I/B/E/S database. The first three columns (permno, ticker and cusip) represent three identifiers for the stock. The variable public_date tells the date of when the company shared its financial statements, i.e. when the financial ratios became public. The columns in Financial ratios (not shown in this table) represent the financial ratios which Apple reported on 30/09/2000. Finally, the integer in the column ffi49 tells in which of the 49 Fama-French industry groups Apple belongs (in this case, it belongs to group 35 which includes companies manufacturing computer hardware). For more details on this industry classification, please refer to French (2020).

| permno | ticker | cusip | public_date | Financial ratios | ffi49 |
|---|---|---|---|---|---|
| 14593 | AAPL | 3783310 | 30/09/2000 | ... | 35 |

### B.4. Macroeconomic Variables

In the original paper, the authors do not fully explain how they transform the datasets containing the values of the macroeconomic variables into usable ones to be fed to the random forest model. Therefore, we outline below the steps which we take to complete this transformation.

Once downloaded, each one of the four macroeconomic variables (consumption, Gross Domestic Product, Industrial Production Index and unemployment rate) presents itself in a table such as the one shown in Figure 1 (this table specifically contains the values of consumption). In this table, the name of the column communicates when the statistics was computed while the name of the row indicates to which period it refers. For example, the entry in the top left corner of this table shows that the estimate for the value of consumption of the fourth quarter (Q4) of 1965 (deducible from the row name) in the fifth month of 1966 (deducible from the last 4 characters of the column title) was 403.3. This value was later revised; indeed, in August 1966 the estimate for the value of consumption for the fourth quarter of 1966 was revised to 406.5 (see 1st row, 4th column). Note that some values are not available. This indicates that the value of the indicator was not yet available for that specific period. For example, in May 1966, the estimate for the value of consumption for the first quarter of 1966 (see the 2nd row, 1st column) was not yet available.

| | RCON66M5 | RCON66M6 | RCON66M7 | RCON66M8 | RCON66M9 |
|---|---|---|---|---|---|
| **1965:Q4** | 403.3 | 403.3 | 403.3 | 406.5 | 406.5 |
| **1966:Q1** | #N/A | 409.9 | 409.9 | 412.8 | 412.8 |
| **1966:Q2** | #N/A | #N/A | #N/A | #N/A | 412.2 |

**Figure 1:** Example of the format of the dataset provided by the Federal Reserve Bank of Philadelphia. These values specifically refer to the US consumption.

The data provided by Van Binsbergen et al. (2023) offers valuable insights, but requires conversion into a time-series format for further analysis. Since the authors don't specify the original time-series structure, we adopt a consistent approach: using the most recently available estimate for each variable at each time step. This methodology, applied to the data in Figure 1, yields the time series depicted in Figure 2. We employ this strategy for all four datasets encompassing the macroeconomic variables.

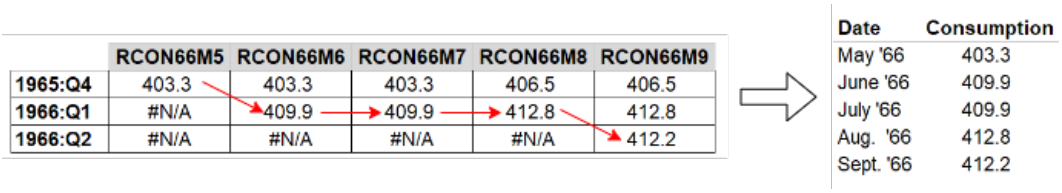



**Figure 2:** Example of how we transform the datasets from the Federal Reserve Bank of Philadelphia into a time-series format.

Finally, we calculate the logarithm of the difference for all macroeconomic time series except the unemployment rate. For example, for the values of consumption, we apply the following transformation:

$$\texttt{Consumption\_log\_return}_t = \ln\left(\frac{\texttt{consumption}_t}{\texttt{consumption}_{t\ 1}}\right) \tag{2}$$

### B.5. Merging and Splitting the Datasets

The final step consists in merging the various datasets described in the previous sections. Since the original paper does not detail the merging process, we adopt the methodology outlined in Figure 3. We believe this approach represents a natural and robust way to combine the data. We take great care to avoid data leakage and to make sure that the merging considered the dates of when the information became public, rather than the end date of the corresponding fiscal period. After merging the datasets into a single one, we split it into five separate datasets based on the forecast horizon.

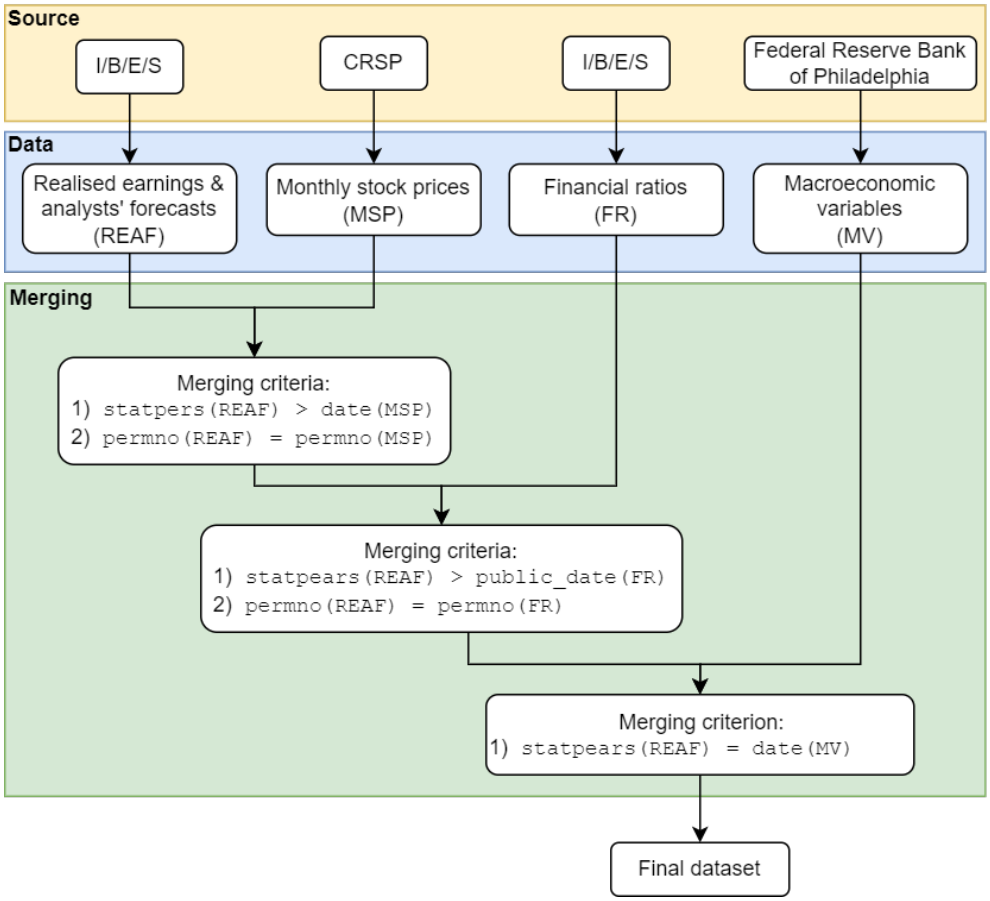



**Figure 3:** Illustration of the merging process. We use the notation `Var(Database)` to indicate that the variable `Var` comes from the database labelled `Database`.

## C. The Model

Following Van Binsbergen et al. (2023)'s approach of leveraging a random forest model to enhance analyst forecasts, we dedicate this section to explain the underlying concepts. Section C.1 delves into the fundamentals of regression trees and random forest regressions. Next, Section C.2 examines the model proposed in the original paper in detail. We unpack its specific functionalities, particularly its role in calculating the conditional bias, a crucial factor to investigate analysts' forecasts.

### C.1. Regression Trees & Random Forest Regressions

Regression trees[3] are a machine learning technique to solve regression problems. They work by recursively partitioning the input space into smaller regions based on specific features. Within each region, the model predicts a constant value for the target variable $y$. This behaviour allows regression trees to capture non-linear relationships between the features and the output.

Let us denote the output of the regression tree for an input vector $x \in \mathbb{R}^K$ as $\hat{y}$. The tree can be described by the following equation:

$$\hat{y} = f_{rt}(x) = \sum_{m=1}^{M} \hat{c}_m I(x \in R_m) \tag{3}$$

where:

- $M$ represents the total number of terminal nodes (or leaves) in the tree.
- $R_m$ denotes the region (defined by a specific set of splitting rules) in the feature space that corresponds to leaf node $m$.
- $\hat{c}_m$ represents the constant value predicted by the model in region $R_m$.
- $I(x \in R_m)$ is the indicator function, which equals 1 if the data point $x$ is in the region $R_m$ and 0 otherwise.

The optimal value of $\hat{c}_m$ in each region depends on the chosen loss function used in the training. By setting the loss function equal to the mean squared error ($MSE$), it can be shown that optimal value of $\hat{c}_m$ becomes the average of the target variable values for the data points within that region:

$$\hat{c}_m = \frac{\sum_{i=1}^{N} y_i I(x \in R_m)}{\sum_{i=1}^{N} I(x \in R_m)} \tag{4}$$

where:

- $y_i$ represents the target variable value for the $i$-th data point.
- $N$ represents the total number of data points in the dataset.

Equation 4 shows that, while calculating the constant value for each region is simple after the split, the key challenge is determining which feature (denoted by $j$) to split on and at what value (denoted by $s$). This split will create two new regions $R_1$ and $R_2$ defined as follows:

$$R_1(j, s) = \{X | X_j \leq s\} \tag{5}$$

$$R_2(j, s) = \{X | X_j > s\} \tag{6}$$

To determine the best split, the algorithm solves an optimization problem: it finds the splitting point $s$ for every feature and then it picks the pair $(j, s)$ which yield the lowest $MSE$. The equation to minimise is shown below:

$$\min_{j,s} \left[ \min_{\hat{c}_1} \sum_{x_i \in R_1(j,s)} (y_i - \hat{c}_1)^2 + \min_{\hat{c}_2} \sum_{x_i \in R_2(j,s)} (y_i - \hat{c}_2)^2 \right] \tag{7}$$

[3]The mathematical treatment of regression trees and random forest regressions is taken from Hastie et al. (2009).

The tree continues to grow until it meets a criterion specified by the user. Common stopping criteria include:

- Minimum Node Size: this criterion stops tree growth when a further split would result in child nodes with fewer data points than a specified threshold.
- Maximum Tree Depth: this criterion limits the maximum depth of a tree, preventing excessive complexity.

The extension from regression trees to random forest is relatively straightforward. The principle behind random forests regressions is to train a lot of approximately unbiased regression trees and then average their predictions to reduce the variance of the model (the bias is, however, unchanged). In order to obtain uncorrelated trees, each tree is trained using the following algorithm:

1. First, select a bootstrap sample of the training data
2. Second, choose $k$ of the $K$ features at random
3. Third, train the random tree using the bootstrapped sample from Step 1 and the variables chosen in Step 2. The tree is then grown until a stopping criterion is met.

The final prediction of a random forest regression model, denoted by $\hat{y}$ or $f_{rf}(x)$, is calculated by averaging the predictions of all the individual trees in the ensemble. Mathematically, this can be expressed as:

$$\hat{y} = f_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{8}$$

where:

- $\hat{y}$ or $f_{rf}(x)$ represents the predicted value for a given input vector $x$.
- $B$ represents the total number of trees in the random forest.
- $T_b(x)$ represents the prediction made by the $b$-th individual tree in the forest when presented with the input vector $x$.

### C.2. Details of the Model

This section focuses on the model employed to forecast adjusted EPS. We delve into its inner workings and how it utilizes the data prepared in the data engineering section as inputs for this prediction task.

Letting $ML\_Forecast_{i,t}^{t+\tau}$ be the random forest forecast of the EPS of company $i$ made at time $t$ to predict the results for the fiscal period ending at time $t+\tau$, where $\tau$ represents the forecast horizon, the regression takes the form of the equation shown below:

$$ML\_Forecast_{i,t}^{t+\tau} = f_{rf}\left(Fundamentals_{i,t}, Macro_t, Analysts\_Forecast_{i,t}^{t+\tau}\right) \tag{9}$$

In this equation, $f_{\mathrm{rf}}(\dots)$ indicates the random forest model which uses the financial ratios ($Fundamentals_{i,t}$) of company $i$, the macroeconomic variables ($Macro_t$), and the average analysts' forecast ($Analysts\_Forecast_{i,t}^{t+\tau}$) for company $i$ available at time $t$ with the same forecast horizon $\tau$ of the model.

One key characteristic of this model is its temporal dependency. In other words, the model is retrained periodically to incorporate the latest information. Specifically, for each forecast horizon, the model is retrained monthly using the most recent 12 months of data. However, for a two-year forecast horizon, the model leverages a larger window of the most recent 24 months of data to capture longer-term trends. Figure 4 provides a visual representation of this rolling window training and testing methodology.

The model's output for the test window is then used to calculate the conditional bias. This metric, denoted by *Conditional_Bias*, reflects the difference between the average analyst forecast and the machine learning prediction, scaled by the company's share price at the time the analyst forecast is made. The mathematical formula for conditional bias is presented below:

$$Conditional\_Bias_{i,t}^{t+\tau} = \frac{Analysts\_Forecast_{i,t}^{t+\tau} - ML\_Forecast_{i,t}^{t+\tau}}{Price_{i,t\ \ 1}} \tag{10}$$

where $Price_{i,t\ \ 1}$ represents the share price of company $i$ at the end of the month before the analysts' average forecast is calculated.

To further illustrate the functionality of the machine learning model, Algorithm 1 presents the training and testing process in pseudocode format.

**Table 5:** Hyper-parameters used for the random forest.

| | |
|---|---|
| Number of trees | 2000 |
| Maximum depth | 7 |
| Sample fraction | 1% |
| Minimum node size | 5 |

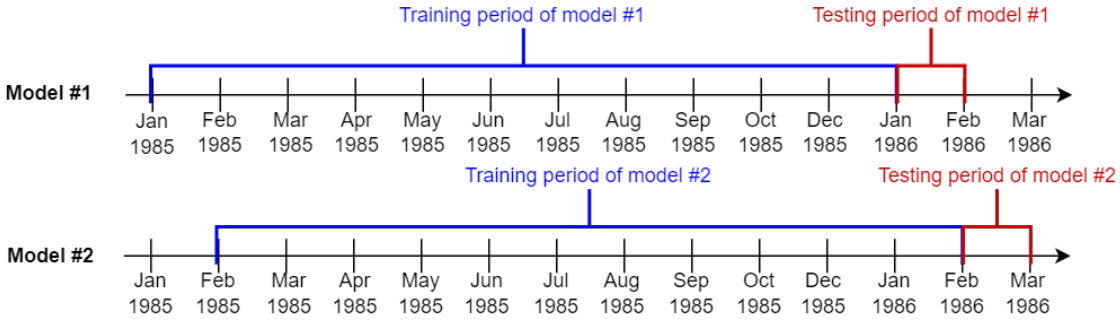



**Figure 4:** Illustration of the *rolling window* method used to train and test the random forest. This diagram shows what are the first two training and test time periods for the models trained using 12 months of data.

**Algorithm 1** Pseudocode of the algorithm used to train and test the model.

**procedure** TrainTestRandomForest($X$, $Y$)
  **for** *forecast_horizon* **in** {3 months, 6 months, 9 months, 12 months, 24 months} **do**
    **if** *forecast_horizon* = 24 months **then**
      $T_{train} \leftarrow$ 24 months ▷ The length of the training period is 24 months for the 2-years forecast horizon
    **else**
      $T_{train} \leftarrow$ 12 months ▷ For all others forecast horizons, the length of the training period is 12 months
    **end if**
    $t_{train} \leftarrow 01/1985$ ▷ Set the beginning of first rolling window to January 1985
    $T_{test} \leftarrow$ 1 month ▷ Set the length of the test period equal to 1 month
    **while** $t_{train} \leq 12/2018$ **do** ▷ While the rolling window is in the correct time-period
      $t_{test} \leftarrow t_{train} + T_{train}$ ▷ Update beginning of test period
      $X_{train} \leftarrow X | t_{train} \leq t < t_{test}$ ▷ Find the input matrix for the training
      $Y_{train} \leftarrow Y | t_{train} \leq t < t_{test}$ ▷ Find the output matrix for the training
      $X_{test} \leftarrow X | t_{test} \leq t < t_{test} + T_{test}$ ▷ Find the input matrix for testing
      $Y_{test} \leftarrow Y | t_{test} \leq t < t_{test} + T_{test}$ ▷ Find the output matrix for the testing
      Train the Random Forest using $X_{train}, Y_{train}$ ▷ Training
      Test the Random Forest using $X_{test}, Y_{test}$ ▷ Testing
      Calculate *Conditional_Bias* ▷ Use the values obtained from the testing phase
      $t_{train} \leftarrow t_{train} + T_{test}$ ▷ Update the beginning of the training window
    **end while**
  **end for**
**end procedure**

## D. Further Study of the Conditional Bias

This section explores the potential drivers of analysts' conditional bias based on observations of the model prediction. We first observe that the conditional bias tends to be lower for stocks with a higher number of analyst coverages, suggesting a richer information environment might lead to more accurate forecasts. Second, the conditional bias appears to be lower on average for the period following the implementation of Regulation Fair Disclosure in October 2000, which prohibits companies from selectively disclosing non-public material information to analysts Reuters (2024).

These observations motivate us to test the following hypotheses:

- The average level of the conditional bias decreases in a statistically meaningful way as the richness of the information environment (measured through the number of analysts covering a particular stock) increases.
- The average level of conditional bias has decreased in a statistically significant way since the introduction of the Regulation on Fair Disclosure in October 2000.

To test these two hypotheses, we fit a linear regression to the values of conditional bias obtained by running Algorithm 1. More specifically, the regression takes the form of the

following equation:

$$Conditional\_Bias_{i,t}^{t+\tau} = b_0\alpha_i + b_1\beta_t + b_2 PR_t + b_3 Analysts\_Number_{i,t} + \epsilon_{i,t} \tag{11}$$

In this equation, the variables $b_0, b_1, b_2, b_3$ represent constant coefficients to be determined. $\alpha_i$ and $\beta_t$ represent, respectively, the time-series average of the conditional bias of firm $i$ and the time-step average of the conditional bias of all firms at time $t$. Mathematically, considering $F$ firms and $T$ measurements conducted at different times of the conditional bias for each firm, the variables $\alpha_i$ and $\beta_t$ may be calculated as follows:

$$\alpha_i = \frac{1}{T}\sum_{t=1}^{T} Conditional\_Bias_{i,t}^{t+\tau} \tag{12}$$

$$\beta_t = \frac{1}{F}\sum_{i=1}^{F} Conditional\_Bias_{i,t}^{t+\tau} \tag{13}$$

The variable $PR_t$ is a dummy variable to study the effect the Fair Disclosure Regulation; this variable is equal to 0 before the end of October 2000 and 1 afterwards. Finally, the variable $Analysts\_Number_{i,t}$ is a positive integer representing the number of analysts covering the stock of firm $i$ at time $t$.

With this regression, we are testing whether the variables representing the Fair Disclosure Regulation and information richness ($PR$ and *Analysts_Number*) can be useful to explain the conditional bias beyond the time-series and firm-specific averages. Thus, to test our hypotheses, we are particularly interested in the values of the coefficients $b_2$ and $b_3$. If both of these coefficients are negative and significant, it would provide evidence to support our hypotheses. It is important to note that we have five separate datasets, one for each forecast horizon considered. Therefore, we conduct the regression described by Equation 11 on the data corresponding to each of these five horizons.

# IV. Exploratory Data Analysis

Building on the Methodology (Section III), this section presents the results obtained from the analysis of the datasets we introduced earlier. We begin by presenting the results of our study of the distribution of earnings in Section A; this study allows us to justify the need to remove outliers from the dataset. Then, in Section B, we attempt to verify the similarity of our dataset to the one of the original paper by comparing their sizes. This comparison reveals that our dataset is smaller and we explain the reasons behind this difference in the same section. Finally, in Section C, we study the relationship between some of the key variables present in the dataset.

## A. Distribution of the EPS Values

Our initial examination of the annual and quarterly EPS dataset revealed a significant number of entries with extreme values. For instance, there are 1,374 quarterly EPS values and 2,266 annual EPS values with an absolute value greater than $10^2$. However, our practical experience teaches us that the values of the EPS tend to be small (usually, in the order of cents for quarterly EPS and in the order of dollars for annual EPS), something which clearly does not align with this finding. Therefore, we decided to investigate further whether these results correspond to some real data or whether they are a wrong entry in the database.

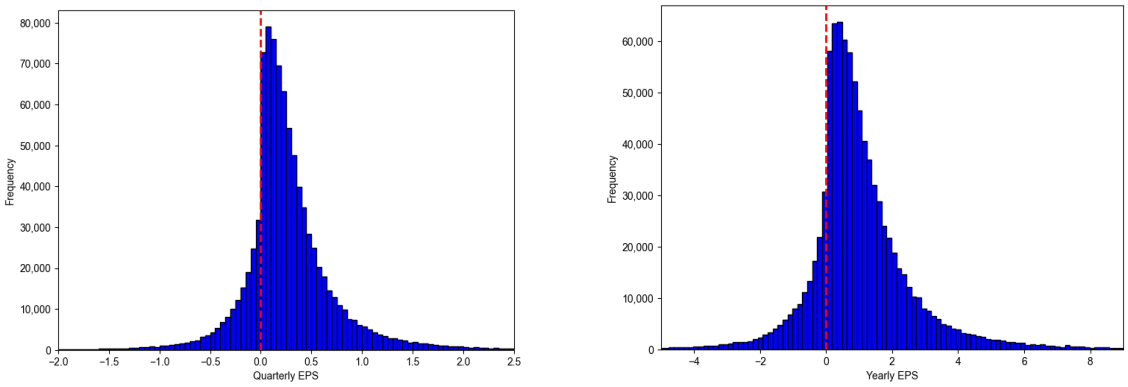
To verify the validity of these potential outliers, we manually compared some of these entries to the EPS values reported on the financial statements available on the U.S. Securities and Exchange Commission (SEC) EDGAR database. Some of these comparisons are shown in Table 6. As it can be seen from this table, some of the EPS values with a large magnitude reported on the I/B/E/S database are wrong. As a result of this finding, we conducted some further research which revealed that the presence of erroneous entries in this database is consistent with the literature on the topic. For example, Ljungqvist et al. (2009) found that, across seven downloads from the I/B/E/S database occurring at an annual frequency between 2000 and 2007, between 1.6% and 21.7% of the entries were different from one download to the other, a result which raised our concerns regarding the validity of the data. This finding was later reproduced by Call et al. (2021), who found a significant difference between two downloads of the I/B/E/S data, with the first one downloaded in 2009, and the second one in 2015. Finally, Acker & Duck (2009) compared 1,874 hand collected data with the entries in the I/B/E/S database and found difference between the reporting dates in 24% of the cases. These studies, together with the results shown in Table 6, led us to believe that the large EPS values previously identified were probably erroneous and should thus be considered as outliers.

**Table 6:** Comparison between of some of the quarterly EPS values obtained from the I/B/E/S database and the corresponding ones obtained from the SEC EDGAR database.

| **Company** | **Ticker** | **Fiscal Period End Date** | **I/B/E/S EPS Value** | **SEC EDGAR EPS Value** |
|---|---|---|---|---|
| Prospect Global | PGRX | 30/06/2013 | -481 | 0.02 |
| Intercloud systems | GRYG | 30/09/2014 | -96 | -0.81 |
| Odonate Therapeutics | ODT | 30/09/2018 | -1960 | -0.98 |

Although random forests are more robust to the presence of outliers than other machine learning models (Hastie et al. 2009), our initial testing reveals that the magnitude of these outliers is so large that the random forest model with the hyper-parameters specified in the original paper is not capable of handling them, causing the model to be incapable of predicting small EPS values. In the original paper, the authors do not specify how they remove these outliers. Therefore, to circumvent the problem, we trim the dataset by removing the 1% outliers. We plot the histograms of the trimmed quarterly and annual EPS values in Figure 5.

One interesting feature which is immediately evident by studying Figure 5 is that the earnings do not follow a symmetrical distribution. Rather, they are quite heavily skewed to the right (in Figure 5, we plot in red a vertical line at EPS = 0 to better show this skew). This skewness is consistent with the literature regarding the distribution of earnings. For instance, Burgstahler & Dichev (1997) find a very similar distribution and explain it via information-processing heuristics and prospect theory. These theories are beyond the scope of this paper; nevertheless, it is reassuring to find this consistency between our observed distribution of earnings and the literature.




**(a)** Quarterly EPS **(b)** Yearly EPS

**Figure 5:** Histograms of the EPS values obtained from the I/B/E/S database.

## B. Size of the dataset

Next, we compare the size of our dataset to the one described in Van Binsbergen et al. (2023). The results of this comparison are presented in Table 7. As it can be seen from this table, the size of our dataset, for all the forecast horizons, is smaller than that of the original paper by a percentage varying from 13.4% to 30.9%. Several factors could potentially explain this discrepancy. First, we refer again to the work Ljungqvist et al. (2009) and Call et al. (2021), who both found that the I/B/E/S dataset can vary depending on the download date. Then, we cannot exclude the possibility that our merging process differed from the one used in Van Binsbergen et al. (2023). The merging process employed in the original paper is not fully described. Since merging can remove unpaired observations and reduce dataset size, a slight difference in their merging procedure could explain the observed size variation. Finally, although we believe that outliers were also removed from the set of target data points in the original paper, their specific methodology for outlier removal remains unknown. This difference in outlier handling could contribute to the size discrepancy.

**Table 7:** Comparison between the size of our dataset and the size of the dataset used in the original paper.

| **Forecast Horizon** | **Number of data points in our dataset** | **Number of data points in the dataset of the original paper** | **Difference** |
|---|---|---|---|
| One quarter | 885,765 | 1,022,661 | -13.4% |
| Two quarters | 811,899 | 1,110,689 | -26.9% |
| Three quarters | 748,251 | 1,018,958 | -26.6% |
| One year | 880,177 | 1,260,060 | -30.1% |
| Two years | 757,719 | 1,097,098 | -30.9% |

## C. Analysis of the Key Variables

We now study the relationships between the key input variables for the random forest model and the target variable (quarterly EPS values). We select these key input variables based on those identified by Van Binsbergen et al. (2023) as having the highest feature importance. More specifically, we focus on:

- Share price (`price`)
- Analysts' average forecast (`meanest`)
- Most recent EPS value (`adj_past_eps`)
- Return on equity over the previous month (`ret`)

To understand these relationships, we employ a pair plot with univariate and bivariate histograms (see Figure 6). This plot is generated using a random sample of 10,000 data points (approximately 1% of the dataset) to ensure visualization clarity.

From the pair plot shown in Figure 6, it is clear that earnings and analysts' predictions show a strong linear relationship. Furthermore, there is a fairly strong correlation between the most recent realisation of earnings and the future ones. This latter finding aligns with the numerous studies detailing the good performance of random walk models to predict EPS values, such as those by Gerakos & Gramacy (2013) and Monahan (2018). Consequently, we expect analysts' forecasts and the latest earnings realisations to hold significant importance in our random forest model.

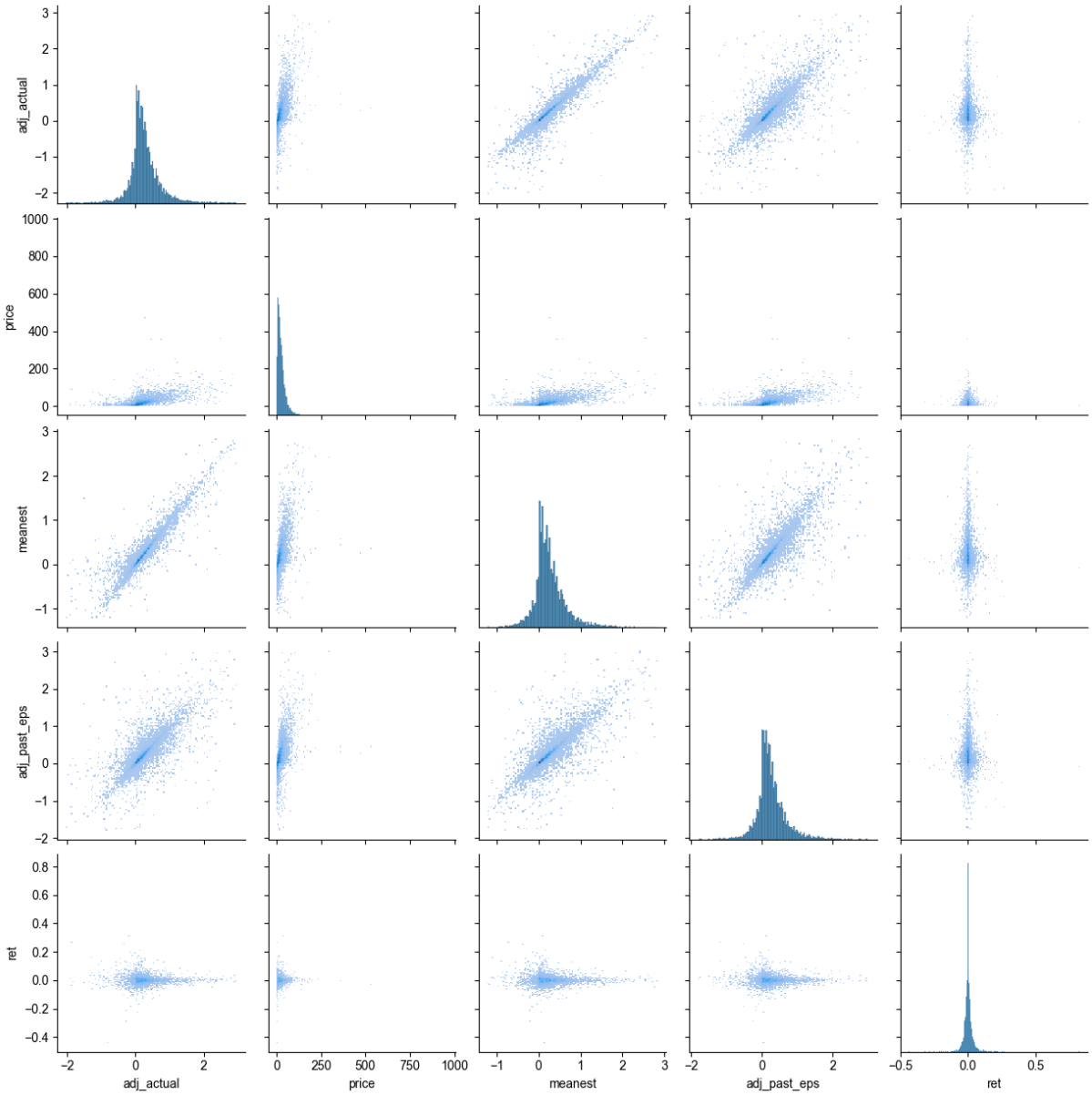



**Figure 6:** Pair plot of the key features identified by looking at the feature importance results of the original paper.

# V. Results

This section analyses the key findings of our work. We begin by presenting, in Section A, the results obtained from the random forest prediction model previously introduced. This section showcases the model's performance metrics, allowing us to assess its effectiveness in predicting future EPS values. In the same section, we compare our results with those of the original paper, demonstrating the success of our model replication. Following this performance evaluation, we explore the concept of conditional bias and its potential drivers. In Section B, we present the results of our analysis of the conditional bias. This analysis examines the relationship between factors such as information richness (measured by the number of analysts covering a stock) and the introduction of the Regulation on Fair Disclosure, and their potential influence on analysts' bias. Our analysis provides supporting evidence that analyst bias diminishes as the information available about a company's stock becomes richer. Additionally, we observe a reduction in bias following the implementation of Regulation Fair Disclosure (FD) by the SEC.

## A. Replication of Original Paper's Findings

This section examines the replication of the original paper's key findings using the random forest model. In Section A.1, we analyse the accuracy of our random forest model and the term structure of the conditional bias. Then, in Section A.2, we compare the feature importance obtained from our analysis with the original paper's findings. Finally, in Section A.3, we analyze and compare the findings on conditional aggregate bias between our study and the original paper. Lastly, in Section A.4 we discuss the key observations and demonstrate that our model successfully replicates Van Binsbergen et al. (2023)'s findings with a high degree of accuracy.

### A.1. Model Performance and Term Structure of the Conditional Bias

To be able to make meaningful comparisons with the original paper's model, we evaluate the out-of-sample performance of our random forest model using the same metrics employed by Van Binsbergen et al. (2023). Table 8 presents the values of these metrics for both our model and the original model. This table facilitates a direct comparison by including both sets of results.

Our results confirm the original paper's finding that analysts tend to be overly optimistic, especially for longer forecasts. As can be seen from Table 8, the average differences between analysts' forecasts and realised earnings (shown in columns labeled $(AF - AE)$) are all positive and increase monotonically as the forecast horizon gets longer, ranging from 0.007 for the one-quarter horizon to 0.216 for the two-year horizon. Furthermore, Figure 7 shows that our values follow a similar trend to the original paper, although they are systematically lower. The observed differences, with an absolute value smaller than 0.1 for all horizons except the two-year horizon, suggest a high similarity between our findings and those of the original paper.

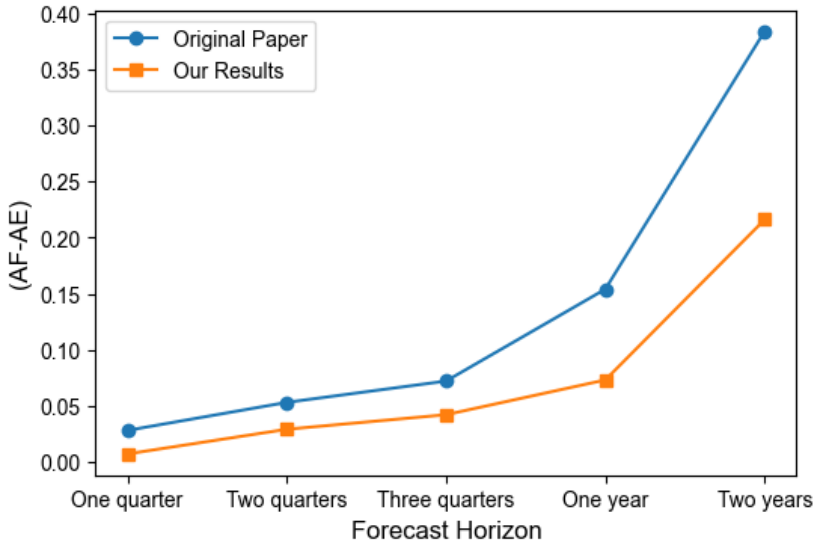



**Figure 7:** Comparison of the values obtained for the term structure of the average difference between analysts' forecasts and realised earnings (we label this average difference $(AF-AE)$ following the notation used in the original paper).

Similar to the original paper, our results show that the time-series average difference between the random forest model's forecasts and realised earnings (columns labeled $(RF-AE)$ in Table 8) is close to zero across all forecast horizons. This indicates strong agreement between our findings and the original paper. Furthermore, the magnitudes of these differences are remarkably similar. The largest absolute value in our results is 0.011, while all others are around 0.001. In the original paper, the largest value was 0.027, with all others similar to ours (around 0.001).

Our results, as evidenced by the lower mean squared error (MSE) for the random forest model compared to the analyst estimates displayed in columns $(RF-AE)^2$ and $(AF-AE)^2$ of Table 8, suggest that the model has the potential to outperform analysts in forecasting accuracy. This finding highlights the model's potential to outperform analysts in forecasting accuracy.

However, we observe a fairly significant difference in the mean squared error for the two-year forecast horizon. Our value deviates significantly (-0.777) from the value reported in the original paper. We suspect this difference might be caused by a different outlier removal methods compared to the one of Van Binsbergen et al. (2023). Since the squared error term amplifies the influence of outliers, this variation could explain the observed discrepancy.

Lastly, considering the term structure of the conditional bias, the original authors find that the average conditional bias is systematically positive and that it increases monotonically with the forecast horizon (from 0.005 for one quarter to 0.021 for two years). They also observe a moderate increase for shorter horizons (one to three quarters) compared to a steeper rise for longer horizons (one and two years). Our findings (see Figure 8 and Table 8, columns labeled $(RF-AF)/P$) mirror this trend. Similar to the original paper, the average conditional bias in our results is consistently positive and exhibits a rising pattern. Additionally, we find identical values (rounded to three decimal places) for the two-quarter and three-quarter horizons. However, our conditional bias values are systematically lower by an average of -0.006 compared to theirs.

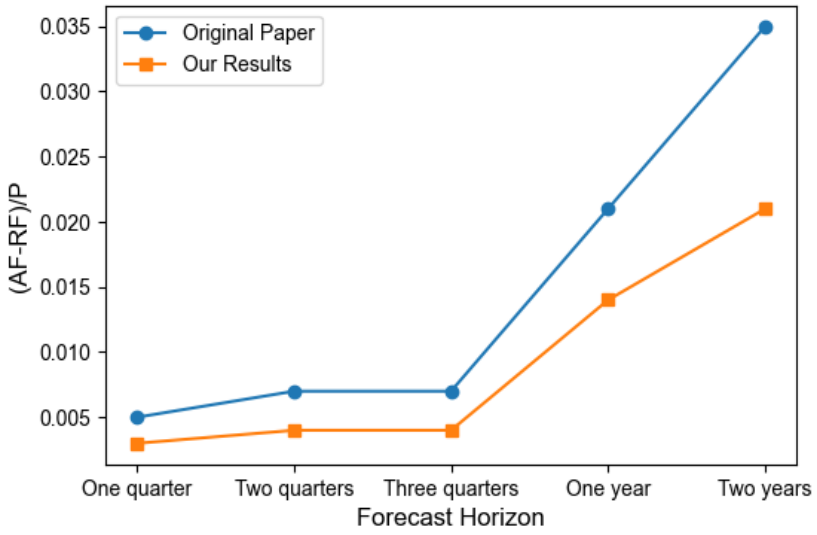



**Figure 8:** Comparison of the term structure of the conditional bias.

### A.2. Feature Importance

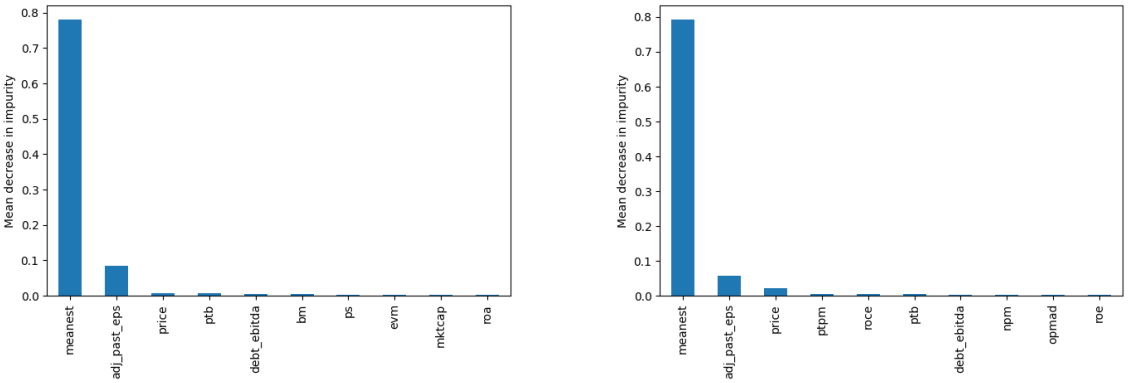
After studying the metrics discussed in Section A.1, Van Binsbergen et al. (2023) examine the importance of the various features used as inputs to the random forest models for the one-quarter and one-year forecasts. In particular, they look at the decrease in impurity by implementing the procedure described by Nembrini et al. (2018). However, in our analysis, we employ a different approach, utilizing scikit-learn's built-in function to calculate the decrease in impurity. Furthermore, since the random forest is retrained monthly, the original paper computes feature importance by averaging the importance values across all models. We replicate this averaging approach in our analysis.

Our feature importance analysis, presented in Figure 9, reveals that we partially agree with Van Binsbergen et al. (2023)'s findings. Similar to the original paper, analysts' forecasts, the most recent EPS, and the most recent share price are the top three most important features for both forecast horizons (one-quarter and one-year).

However, we diverge in the importance of return on capital employed (ROCE) and return on equity (ROE). Unlike the original paper, where these metrics ranked fourth or fifth for both horizons, our results show them only reaching the top 10 for the one-year forecast model. Finally, consistent with Van Binsbergen et al. (2023), all remaining features exhibit low importance.

**Table 8:** Comparison of the numerical results obtained from our replication of the original paper with the ones presented in Table 2 of the original paper. Columns labelled *RF* show the average value of the forecasts made by the random forest. Columns labelled *AF* present the average value of analysts' forecasts, while those labelled *AE* display the average value of the realised earnings. Columns titled *(RF-AE)* show the average difference between the random forest's forecasts and the realised earnings, while those labelled *(AF-AE)* present the average difference between analysts' forecasts and the realised earnings. Columns under the title $(RF - AE)^2$ show the average of the squared difference between the random forest's predictions and the analysts' forecasts. Similarly, columns labelled $(AF - AE)^2$ show the average of the squared difference between the analysts' forecasts and the realised earnings. Finally, columns labelled $(AF - AP)/P$ show the average value of the conditional bias which is calculated using Equation 10.

| | **RF** | | **AF** | | **AE** | | **(RF-AE)** | | **(AF-AE)** | | **$(RF-AE)^2$** | | **$(AF-AE)^2$** | | **(AF-RF)/P** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Forecast Horizon | Original Paper | Our Result | Original Paper | Our Result | Original Paper | Our Result | Original Paper | Our Result | Original Paper | Our Result | Original Paper | Our Result | Original Paper | Our Result | Original Paper | Our Result |
| One quarter | 0.290 | 0.259 | 0.319 | 0.267 | 0.291 | 0.260 | 0.000 | −0.001 | 0.028 | 0.007 | 0.076 | 0.034 | 0.081 | 0.042 | 0.005 | 0.003 |
| *Difference* | | −0.031 | | −0.052 | | −0.031 | | −0.001 | | −0.021 | | −0.042 | | −0.039 | | −0.002 |
| Two quarters | 0.323 | 0.276 | 0.376 | 0.305 | 0.323 | 0.276 | −0.001 | 0.000 | 0.053 | 0.029 | 0.094 | 0.050 | 0.102 | 0.061 | 0.007 | 0.004 |
| *Difference* | | −0.047 | | −0.071 | | −0.047 | | 0.001 | | −0.024 | | −0.044 | | −0.041 | | −0.003 |
| Three quarters | 0.343 | 0.296 | 0.413 | 0.337 | 0.341 | 0.295 | 0.002 | 0.001 | 0.072 | 0.042 | 0.121 | 0.066 | 0.132 | 0.088 | 0.007 | 0.004 |
| *Difference* | | −0.047 | | −0.076 | | −0.046 | | −0.001 | | −0.030 | | −0.055 | | −0.044 | | −0.003 |
| 1 year | 1.194 | 1.045 | 1.32 | 1.121 | 1.167 | 1.049 | 0.027 | −0.003 | 0.154 | 0.073 | 0.670 | 0.404 | 0.686 | 0.433 | 0.021 | 0.014 |
| *Difference* | | −0.149 | | −0.199 | | −0.118 | | −0.030 | | −0.081 | | −0.266 | | −0.253 | | −0.007 |
| 2 years | 1.384 | 1.263 | 1.771 | 1.468 | 1.387 | 1.252 | −0.004 | 0.011 | 0.384 | 0.216 | 1.897 | 1.120 | 2.009 | 1.747 | 0.035 | 0.021 |
| *Difference* | | −0.121 | | −0.303 | | −0.135 | | 0.015 | | −0.168 | | −0.777 | | −0.262 | | −0.014 |
| *Average difference* | | −0.079 | | −0.140 | | −0.075 | | −0.003 | | −0.065 | | −0.237 | | −0.128 | | −0.006 |




**(a)** One-quarter forecast horizon. **(b)** One-year forecast horizon.

**Figure 9:** Histograms of the feature importance for the 10 most important features in the one-quarter and one-year forecast horizon machine learning models.

### A.3. Conditional Aggregate Bias

Finally, Van Binsbergen et al. (2023) also examine the conditional aggregate bias, which they define as the average of the individual stocks' conditional bias at each time-step. They first study the conditional aggregate bias relative to the machine learning forecasts, and then the conditional aggregate bias relative to the realised earnings. We repeat this analysis following the same methodology; our results are presented in Figure 10 and 11.

Analysing these two figures, we observe that they look qualitatively very similar to the ones presented in the original paper. For instance, Van Binsbergen et al. (2023) also find clear spikes during the periods corresponding to the Internet bubble and the 2007-2008 financial crisis. Furthermore, similarly our plots, the plots in the original paper also show that, on average, at every time-step as the forecast horizon increases, the conditional aggregate bias also increases. Also, similarly to their values, our values in both figures are almost always positive. Finally, when comparing the heights of our peaks to the ones of the plots in the original paper, we observe that our peaks are consistently lower by approximately 0.015.

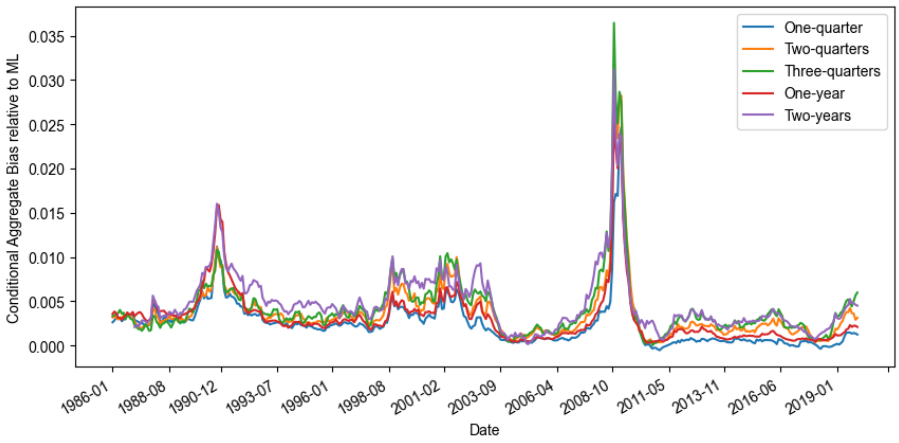



**Figure 10:** Conditional aggregate bias relative to the machine learning forecasts.

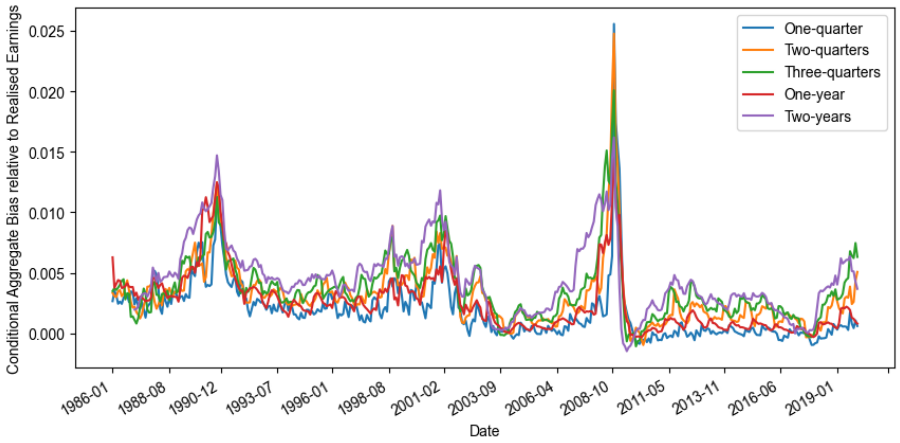



**Figure 11:** Conditional aggregate bias relative to the realised earnings.

## A.4. Discussion of Replication Results

In this section, we discuss the results obtained from the replication of the original paper. In particular, we present the notion that our results suggest that the findings of the original paper can be replicated. We make this claim based on several observations.

First, we obtain performance metrics that are very similar to those calculated in the original paper and follow similar trends. Although our values for these metrics are systematically lower than those shown in the original paper (this is particularly evident in the row labelled *Average difference* of Table 8), the difference is almost always very small. The only case where the difference is large is when we compare the mean squared error of the random forest model. We believe that this difference is caused by the fact that we used a different outlier removal algorithm from the original authors, as the paper does not specify which one they used. The presence of these outliers is then most heavily seen in the mean squared error of the model, due to the quadratic nature of these values.

Second, considering the results obtained from the feature importance analysis, we find that the three most important features are the same ones with the highest importance in the original authors' models. We also find that other variables with high predictive power in the original paper also have high importance in our models, although they are ranked in a slightly different order from those shown in the original paper. We hypothesise that there could be two main reasons for this difference. The first reasons is the already mentioned fact that there is a possibility that we use a different outlier removal algorithm from the original authors. The second reason is the fact that we measure feature importance with an algorithm slightly different from the one used by the original authors.

Third, considering the plots of the conditional aggregate bias (both the one relative to the machine learning algorithm as well as the one relative to the realized earnings, which we show in Figure 10 and 11), we find that these plots look qualitatively very similar to those presented in the original paper. Specifically, although our values are, similar to the performance metrics presented in Table 8, slightly lower than those in the original paper, we observe the same patterns, namely the presence of peaks at the same time steps and a general decrease in the level of conditional bias after 2001. Consequently, based on all these observations, we suggest that our replication of the original paper can be considered successful.

## B. Novel Results on Analyst Bias

In this section, we firstly illustrate the observations which motivate our further analysis on the bias. The first observation we make regards the time-series plot of the conditional aggregate bias relative to the machine learning forecasts shown in Figure 10. Studying this plot, one may observe that, before 2001, all the time-series appear incapable of falling below an imaginary horizontal barrier, situated at roughly 0.0025. On the other hand, after 2001, this imaginary barrier appears to be closer to zero. Recalling that the Regulation on Fair Disclosure came into effect in October 2000 Reuters (2024), this observation motivates research into the effect of this new regulation on analysts' conditional bias.

The second observation which we make is regarding the number of analysts covering each stock. When plotting a graph of the average conditional bias as a function of the number of analysts, we observe that, as the number of analysts increases, the average bias tends to decrease. An example of such a plot is shown in Figure 12; as can be seen from this plot, as the number of analysts covering a stock increases, the conditional bias initially decreases rapidly, then approaches zero asymptotically. Although the values slightly change depending on the forecast horizon, the shape of the plot remains consistent. Note that, Figure 12 displays a single data point for each number of analysts, representing the average conditional bias across all firms and all time periods.

To further explore this connection between the number of analysts and the conditional bias, Figure 13 presents the average conditional bias of analysts' earnings expectations relative to the machine learning forecast, grouped by the number of analysts (quartiles) for the one-year horizon. Note that it shows the time series of the average conditional bias for each quartile of analyst coverage.

As expected from Figure 12, Figure 13 visually confirms a systematic relationship. The time series corresponding to the first quartile (fewest analysts) consistently exhibits the highest conditional bias, followed by the second, third, and fourth quartiles (increasing analyst coverage), respectively. This finding reinforces the notion that a larger number of analyst participating to the forecast tends to mitigate the overall conditional bias.

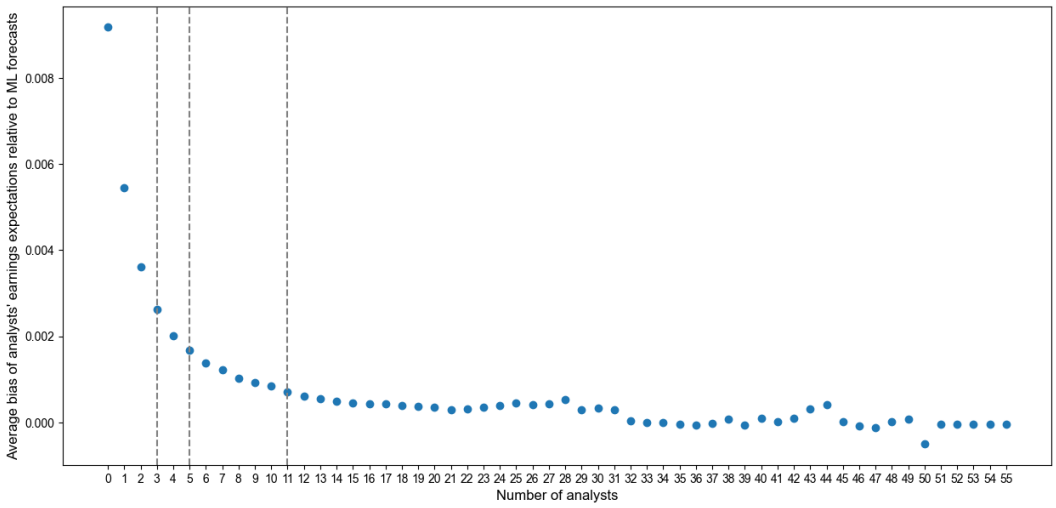



**Figure 12:** Conditional aggregate bias (relative to the machine learning forecasts) vs number of analysts for the one-year forecast horizon. The vertical grey lines indicates the quartiles of the number of analysts.

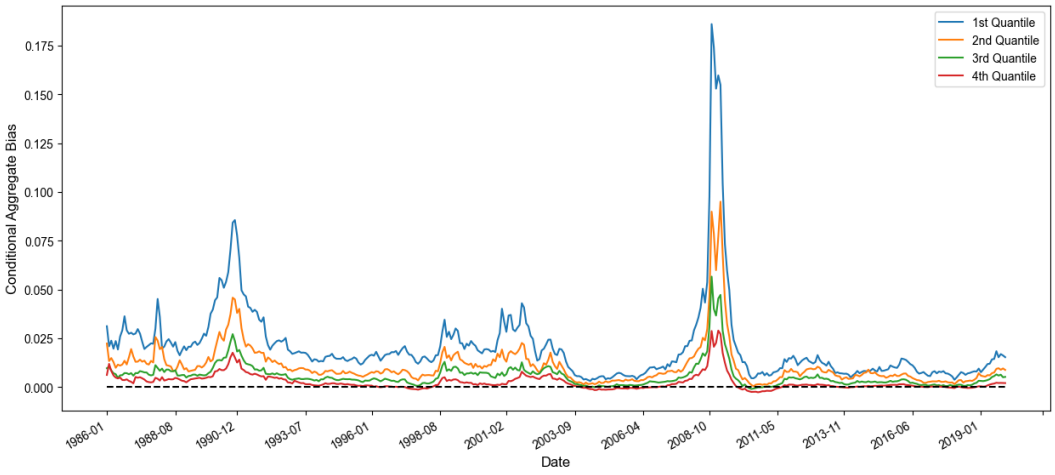



**Figure 13:** Average conditional bias of analysts' earnings expectations relative to machine learning forecast, grouped by number of analysts (quartiles) for the one-year forecast horizon.

### B.1. Statistical Analysis of Results

Building upon the observations from the previous section, we investigate deeper to statistically analyse the factors influencing conditional bias. Here, we employ our methodology described in the Methodology section to assess the impact of various variables on the bias.

The results of our regression analyses, presented in Table 9 (one model for each forecast horizon), reveal several key findings. Firstly, all estimated coefficients across the regressions achieve statistical significance at the 5% level. This signifies a strong association between the independent variables and conditional bias.

Second, the coefficients associated with the firm-specific effect ($\alpha_i$, representing the time-series average conditional bias for each firm) are all positive and very close to 1. These coefficients decrease slightly with longer forecast horizons. This suggests a strong positive correlation between a firm's historical conditional bias and its future conditional bias.

Third, the coefficients for the time-specific effect ($\beta_t$, representing the average conditional bias for all firms at a specific time) are also positive and exhibited a similar trend of decreasing values with longer forecast horizons.

However, the most interesting results are those regarding the coefficients in front of the $PR_t$ and $Analysts\_Number_{i,t}$ terms (representing, respectively, the dummy variable for the introduction of the Fair Disclosure Regulation and the number of analysts). We find these coefficients to be small and negative for all time horizons. Furthermore, we observe that both these coefficients become more negative as the forecast horizon increases.

**Table 9:** Coefficients obtained from the linear regression on the conditional bias versus the four explanatory variables identified in the previous section are reported below. Below each coefficient, we report the corresponding $t$-statistic.

| Forecast Horizon | $\alpha_i$ | $\beta_t$ | $PR_t$ | $Analysts\_Number_{i,t}$ |
|---|---|---|---|---|
| One-quarter | 0.9977 | 0.6662 | -0.0003 | -0.0001 |
| | (889.35) | (54.36) | (-3.56) | (-16.08) |
| Two-quarters | 0.9958 | 0.6703 | -0.0011 | -0.0002 |
| | (527.91) | (44.67) | (-7.32) | (-11.37) |
| Three-quarters | 0.9977 | 0.6253 | -0.0014 | -0.0002 |
| | (718.48) | (41.44) | (-8.60) | (-9.76) |
| One-year | 0.9747 | 0.6141 | -0.0020 | -0.0006 |
| | (451.96) | (79.88) | (-8.60) | (-32.41) |
| Two-years | 0.9663 | 0.5272 | -0.0033 | -0.0007 |
| | (1002.02) | (171.15) | (-29.06) | (-74.98) |

### B.2. Financial Interpretation of the Novel Results

Considering the impact of the Fair Disclosure Regulation, we find that, for all forecast horizons, the conditional bias decreases after the introduction of this regulation. This finding is supported by the fact that, for all forecast horizons, the coefficients in the studied regression in front of the dummy variable $PR_t$ are negative and significant. To explain this finding, we refer to the work of Lim (2001). In this paper, the author presents a model to describe analysts' tradeoff between better access to management information and forecast accuracy. According to this model, analysts have to choose between making more optimistic forecasts, which allow them to have better access to information (used to improve future forecast accuracy), or making more pessimistic (but perhaps accurate) forecasts, which in turn would lead to worse access to information and potentially more inaccurate forecasts in the future.

According to Lim (2001)'s model, supported by empirical evidence, it is rational for analysts to exhibit a positive bias to gain better access to information. We believe that this explanation can also be used to explain our findings. Specifically, before the introduction of the Regulation on Fair Disclosure, companies could disclose material non-public information, which could be used to significantly improve forecasting accuracy. However, after the implementation of this regulation, companies can no longer distribute material non-public information to analysts, meaning that the potential improvement gained by issuing optimistic forecasts has now been reduced. Thus, we suggest that the introduction of this regulation reduced the rational incentives for analysts to make forecasts with a positive bias.

Considering the fact that the decrease in conditional bias after the introduction of the Fair Disclosure Regulation was larger for longer forecast horizons, we believe this relates to the time-compounding effect of changes in growth rates. In other words, if analysts suddenly become less optimistic in their forecasts, decreasing the growth rates, the forecasts with longer horizons will be more impacted than those with a short forecasting horizon.

Moving on to the study of the impact of the number of analysts on the conditional bias, our results suggest that there is a negative and significant relationship between the number of analysts covering a particular stock and the conditional bias. We interpret the number of analysts as a proxy for the richness of the information set regarding a stock, as each analyst contributes to this information set with their analysis. Therefore, we find that as the

information set of a company's stock expands, the conditional bias decreases. We believe that this finding may be explained by the work of Das et al. (1998), who suggest that analysts' bias tends to be more positive for companies with poor information sets. In such cases, the marginal improvement in forecasting accuracy resulting from better access to companies' non-public, immaterial information is more bigger. In other words, there is a higher demand for non-public information in these companies. To access this information, analysts issue more positive forecasts to please management, explaining the negative relationship between conditional bias and the number of analysts. As for the observation that the conditional bias is more strongly related to the number of analysts for longer forecast horizons, we refer again to the same explanation used to describe the impact of regulation, namely, the time-compounding effect of changes in growth rates.

# VI. Conclusions

This dissertation investigated the phenomenon of analyst bias in earnings forecasts and its impact on equity valuations. Accurate earnings forecasts are crucial for efficient capital allocation, yet analysts are known to exhibit a positive bias towards optimism.

Building upon the work of Van Binsbergen et al. (2023), we successfully replicated their real-time measure of conditional bias, confirming its effectiveness in quantifying this bias. Our results, largely aligning with the original paper, demonstrate that the measure is replicable and analysts tend to overestimate future earnings.

Expanding upon our initial analysis, we explored the factors contributing to this bias. Two key determinants emerged: the introduction of Regulation Fair Disclosure in 2000 and the number of analysts following a particular stock. We link these findings to existing literature on the topic discussing bias. Our findings suggest that this regulation, by prohibiting disclosure of material information, made analysts have less incentive to gain management favor to access such information. Furthermore, we observed an inverse relationship between the number of analysts and conditional bias. This observation may be used to suggest that, as the number of analysts grows, the demand for non-public information decreases, meaning that analysts are less incentivised to issue optimistic forecasts to gain better access to management information.

These findings hold significant implications for corporate finance. The measure of conditional bias allows for more accurate stock valuations, potentially leading to a more efficient allocation of capital. Furthermore, our study highlights the importance of regulations such as the Regulation Fair Disclosure in mitigating analyst bias. Finally, the negative association between analyst coverage and bias underscores the vital role analysts play in fostering market efficiency through their collective effort.

# References


Acker, D. & Duck, N. W. (2009), 'On the reliability of i/b/e/s earnings announcement dates and forecasts'.

Albrecht, W. S., Lookabill, L. L. & McKeown, J. C. (1977), 'The time-series properties of annual earnings', *Journal of Accounting Research* pp. 226–244.

Ball, R. & Watts, R. (1972), 'Some time series properties of accounting income', *The Journal of Finance* **27**(3), 663–681.

Beyer, A. & Guttman, I. (2011), 'The effect of trading volume on analysts' forecast bias', *The Accounting Review* **86**(2), 451–481.

Bradshaw, M. T., Drake, M. S., Myers, J. N. & Myers, L. A. (2012), 'A re-examination of analysts' superiority over time-series forecasts of annual earnings', *Review of Accounting Studies* **17**, 944–968.

Burgstahler, D. & Dichev, I. (1997), 'Earnings management to avoid earnings decreases and losses', *Journal of accounting and economics* **24**(1), 99–126.

Call, A. C., Hewitt, M., Watkins, J. & Yohn, T. L. (2021), 'Analysts' annual earnings forecasts and changes to the i/b/e/s database', *Review of Accounting Studies* **26**, 1–36.

Callen, J. L., Kwan, C. C., Yip, P. C. & Yuan, Y. (1996), 'Neural network forecasting of quarterly accounting earnings', *International journal of forecasting* **12**(4), 475–482.

Cao, Q. & Parry, M. E. (2009), 'Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm', *Decision Support Systems* **47**(1), 32–41.

Cao, S., Jiang, W., Wang, J. & Yang, B. (2021), 'From man vs. machine to man machine: The art and ai of stock analyses', *Journal of Financial Economics forthcoming; Columbia Business School Research Paper* .

Dai, R. (2020), 'I/b/e/s @wrds 101'. Accessed: 03/06/2024.
**URL:** *https://wrds-www.wharton.upenn.edu/documents/1372/IBES_RUI.pptx*

Das, S., Levine, C. B. & Sivaramakrishnan, K. (1998), 'Earnings predictability and bias in analysts' earnings forecasts', *Accounting Review* pp. 277–294.

De Silva, T. & Thesmar, D. (2024), 'Noise in expectations: Evidence from analyst forecasts', *The Review of Financial Studies* **37**(5), 1494–1537.

Fama, E. F. & French, K. R. (2000), 'Forecasting profitability and earnings', *The journal of business* **73**(2), 161–175.

French, K. R. (2020), 'Detail for 49 industry portfolios'. Accessed: 03/06/2024.
**URL:** *https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_49_ind_port.html*

Gerakos, J. & Gramacy, R. (2013), 'Regression-based earnings forecasts', *Chicago Booth Research Paper* **26**(12).

Green, J. & Zhao, W. (2022), 'Forecasting earnings and returns: A review of recent advancements', *The Journal of Finance and Data Science* **8**, 120–137.

Gu, Z. & Wu, J. S. (2003), 'Earnings skewness and analyst forecast bias', *Journal of Accounting and Economics* **35**(1), 5–29.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer.

Hess, D. & Wolf, S. (2014), 'Quarterly earnings information: Implications for annual earnings forecast models', *Available at SSRN* .

Hou, K., Van Dijk, M. A. & Zhang, Y. (2012), 'The implied cost of capital: A new approach', *Journal of Accounting and Economics* **53**(3), 504–526.

Jarrett, J. (1989), 'Forecasting monthly earnings per share—time series models', *Omega* **17**(1), 37–44.

Johnson, T. E. & Schmitt, T. G. (1974), 'Effectiveness of earnings per share forecasts', *Financial Management* pp. 64–72.

Koller, T., Goedhart, M. & Wessels, D. (2010), *Valuation - Measuring and Managing the Value of Companies*, 5 edn, John Wiley & Sons, New Jersey.

Koller, T., Goedhart, M., Wessels, D. et al. (2010), *Valuation: measuring and managing the value of companies*, Vol. 499, john Wiley and sons.

Kothari, S. P., So, E. & Verdi, R. (2016), 'Analysts' forecasts and asset pricing: A survey', *Annual Review of Financial Economics* **8**, 197–219.

Li, K. K. & Mohanram, P. (2014), 'Evaluating cross-sectional forecasting models for implied cost of capital', *Review of Accounting Studies* **19**, 1152–1185.

Lim, T. (2001), 'Rationality and analysts' forecast bias', *The journal of Finance* **56**(1), 369–385.

Lipe, R. & Kormendi, R. (1994), 'Mean reversion in annual earnings and its implications for security valuation', *Review of quantitative finance and accounting* **4**, 27–46.

Ljungqvist, A., Malloy, C. & Marston, F. (2009), 'Rewriting history', *The Journal of Finance* **64**(4), 1935–1960.

Monahan, S. (2018), *Financial Statement Analysis and Earnings Forecasting*, Foundations and trends in accounting, Now Publishers.
**URL:** *https://books.google.it/books?id=5w1jzgEACAAJ*

Nembrini, S., König, I. R. & Wright, M. N. (2018), 'The revival of the gini importance?', *Bioinformatics* **34**(21), 3711–3718.

Reuters, T. (2024), 'Regulation fd'. Accessed: 03/06/2024.
**URL:** *https://uk.practicallaw.thomsonreuters.com/7-107-7127*

Scherrmann, M. & Elsas, R. (2023), 'Earnings prediction using recurrent neural networks', *Arxiv* .

Van Binsbergen, J. H., Han, X. & Lopez-Lira, A. (2023), 'Man versus machine learning: The term structure of earnings expectations and conditional biases', *The Review of financial studies* **36**(6), 2361–2396.

Watts, R. L. & Leftwich, R. W. (1977), 'The time series of annual accounting earnings', *Journal of Accounting Research* pp. 253–271.

Xinyue, C., Zhaoyu, X. & Yue, Z. (2020), 'Using machine learning to forecast future earnings', *Atlantic Economic Journal* **48**, 543–545.

Yu, H., Hao, X., Wu, L., Zhao, Y. & Wang, Y. (2023), 'Eye in outer space: satellite imageries of container ports can predict world stock returns', *Humanities and Social Sciences Communications* **10**(1), 1–16.

Zhang, W., Cao, Q. & Schniederjans, M. J. (2004), 'Neural network earnings per share forecasting models: A comparative analysis of alternative methods', *Decision Sciences* **35**(2), 205–237.