Panel Data Nowcasting: The Case of Price-Earnings Ratios*

Andrii Babii† Ryan T. Ball‡ Eric Ghysels§ Jonas Striaukas¶

July 23, 2023

Abstract

The paper uses structured machine learning regressions for nowcasting with panel data consisting of series sampled at different frequencies. Motivated by the problem of predicting corporate earnings for a large cross-section of firms with macroeconomic, financial, and news time series sampled at different frequencies, we focus on the sparse-group LASSO regularization which can take advantage of the mixed frequency time series panel data structures. Our empirical results show the superior performance of our machine learning panel data regression models over analysts' predictions, forecast combinations, firm-specific time series regression models, and standard machine learning methods.

Keywords: Corporate earnings, nowcasting, data-rich environment, high-dimensional panels, mixed frequency data, textual news data, sparse-group LASSO.

---

*We benefited from comments by Rudy De Winne, Geert D'Haene, Max Farrell, Christian Hafner, Peter Reinhard Hansen, Dacheng Xiu, and participants at the 2021 SoFiE UC San Diego conference, 26th International Panel Data Conference, Data Science and Machine Learning workshop at the University of Amsterdam, the 2022 IAAE Conference, King's College, London, and the 2022 Vienna Copenhagen Conference on Financial Econometrics. This work was in part completed when Jonas Striaukas was a Research Fellow at Fonds de la Recherche Scientifique FNRS.

†University of North Carolina at Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305. Email: babii.andrii@gmail.com.

‡Stephen M. Ross School of Business, University of Michigan, 701 Tappan Street, Ann Arbor, MI 48109. Email: rtball@umich.edu.

§Department of Economics and Kenan-Flagler Business School, University of North Carolina–Chapel Hill. Email: eghysels@unc.edu.

¶Department of Finance, Copenhagen Business School, Frederiksberg, Denmark. Email: jonas.striaukas@gmail.com.

1 Introduction

Nowcasting is intrinsically a mixed frequency data problem as the object of interest is a low-frequency data series — observed say quarterly — whereas real-time information — daily, weekly or monthly — during the quarter can be used to assess and potentially continuously update the state of the low-frequency series, or put differently, *nowcast* the series of interest. Traditional methods being used for nowcasting rely on dynamic factor models which treat the underlying low-frequency series of interest as a latent process with high-frequency data noisy observations. These models are naturally cast in a state-space form, and inference can be performed using standard techniques (in particular the Kalman filter, see Bańbura, Giannone, Modugno, and Reichlin (2013) for a recent survey).

Things get more complicated when we are operating in a data-rich environment *and* we have many target variables. Put differently, we are no longer interested in nowcasting a single key series such as the GDP growth where we could devote a lot of resources to that particular series. A good example is corporate earnings nowcasting for a large cross-section of corporate firms. The fundamental value of equity shares is determined by the discounted value of future payoffs. Every quarter investors get a glimpse of firms' potential payoffs with the release of corporate earnings reports. In a data-rich environment, stock analysts have many indicators regarding future earnings that are available much more frequently. Ball and Ghysels (2018) took a first stab at automating the process using MIDAS regressions. Since their original work, much progress has been made on machine learning (ML) regularized mixed frequency regression models.

In the context of earnings, we are potentially dealing with a large set of individual firms for which there are many predictors. From a practical point of view, this is clearly beyond the realm of nowcasting using state space models. In the current paper, we significantly expand the tools of nowcasting in a data-rich environment by exploiting panel data structures. Panel data regression models are well suited for the firm-level data analysis as both the time series and cross-sectional dimensions can be exploited. In such models, time-invariant firm-specific effects are typically used to capture cross-sectional heterogeneity in the data. This is combined with regularized regression machine learning methods which are becoming increasingly popular in economics and finance as a flexible way to model predictive relationships via variable selection. We focus on the panel data regressions in a high-dimensional data setting where the number of covariates could be large and potentially exceed the available sample size. This may happen when the number of firm-specific characteristics, such as textual analysis news data or firm-level stock returns, is large, and/or the number

of aggregates, such as market returns, macro data, etc., is large.

Our paper relates to several existing papers in the literature. Khalaf, Kichian, Saunders, and Voia (2021) consider low-dimensional dynamic mixed frequency panel data models but do not deal with high-dimensional data situations in the context of nowcasting or forecasting. Similarly, Fosten and Greenaway-McGrevy (2019) consider nowcasting with a mixed-frequency VAR panel data model, but not in the context of a high-dimensional data-rich environment that we are interested in here. Babii, Ball, Ghysels, and Striaukas (2022) introduce the sparse-group LASSO (sg-LASSO) regularization machine learning methods for heavy-tailed dependent panel data regressions potentially sampled at different time series frequencies. They derive oracle inequalities for the pooled and fixed effects models, the debiased inference for pooled regression, and consider an application to the Granger causality testing. In this paper, we explore how to use their framework for nowcasting large panels of low-frequency time series.

We focus on nowcasting current quarter firm-specific price-earnings ratios (henceforth P/E ratios). This means we focus on evaluating model-based within-quarter predictions for very short horizons. It is widely acknowledged that P/E ratios are a good indicator of the future performance of a company and, therefore, are used by analysts and investment professionals to base their decisions on which stocks to pick for their investment portfolios. Typically investors rely on consensus forecasts of earnings made by a pool of analysts. We, therefore, choose such consensus forecasts as the benchmark for our proposed machine learning methods. Ball and Ghysels (2018) and Carabias (2018) documented that analysts tend to focus on their firm/industry when making earnings predictions while not fully taking into account the impact of macroeconomic events. Babii, Ball, Ghysels, and Striaukas (2022) tested formally in a high-dimensional data setting the hypothesis that systematic and predictable errors occur in analyst forecasts and confirmed empirically that they *leave money on the table*. The analysis in the current paper is therefore an logical extension of this prior work. In addition, we also compare our proposed new methods with the MIDAS regression forecast combination approach used by Ball and Ghysels (2018) as well as a simple random walk model.

Our high-frequency regressors include traditional macro and financial series as well as non-standard series generated by textual analysis of financial news. We consider structured pooled and fixed effects sg-LASSO panel data regressions with mixed frequency data (sg-LASSO MIDAS). By "structured" we mean that the ML procedure is set up such that it recognizes the time series and panel structure of the data. This is a departure from standard ML which is rooted in a tradition of i.i.d.

covariates and therefore time series and panel data structures are not recognized. For the purpose of comparison, we include elastic net estimators in our analysis, as a representative example of standard ML.

In our empirical analysis we study nowcasting the firm-level P/E ratio for a large set of firms. Moreover, we decompose the (log of) the P/E ratio into the return for firm i and analyst prediction errors. Therefore, nowcasting the log P/E ratio could also be achieved via nowcasting its two components. The decomposition corresponds to the distinction between analyst assessments of firm i's earnings and market/investor assessments of the firm.

Our empirical results can be summarized as follows. Predictions based on analyst consensus exhibit significantly higher mean squared forecast errors (MSEs) compared to model-based predictions. These model-based predictions involve either direct log P/E ratio nowcasts or their individual components. The MSE for the random walk model and analysts' concensus are quite similar, and therefore random walk predictions are outperformed by the model-based ones as well. A substantial proportion of firms (approximately 60%) exhibit low MSE values, indicating a high level of prediction accuracy. However, there are a few firms for which the MSEs are relatively larger, suggesting lower prediction performance for these specific cases. Comparing direct log P/E ratio nowcasts versus those based on its components, we observe a substantial improvement in prediction accuracy when using the individual components. This improvement is consistently evident across individual, pooled, and fixed effects regression models. Moreover, the sparsity patterns differ significantly across the direct versus component prediction models.

Our framework allows us to go beyond providing quarterly nowcasts and generate daily updates of earnings series. Leveraging the daily influx of information throughout the quarter, we continuously re-estimate our models and produce nowcast updates as soon as new data becomes available. We report the distribution of Mean Squared Errors (MSEs) across firms for five distinct nowcast horizons: 20-day, 15-day, 10-day, and 5-day ahead, as well as the end of the quarter and show that as the horizons become shorter, both the median and upper quartile of MSEs decrease. The sg-LASSO estimator we employ in our study is well-suited for incorporating grouped fixed effects. This approach involves grouping firm-specific intercepts based on either statistical procedures or economic reasoning, as outlined in Bonhomme and Manresa (2015). In our analysis, we utilize the Fama French industry classification to form 10 distinct groups for grouping fixed effects. Our findings suggest that grouped fixed effects strike a better balance between capturing heterogeneity and pooled parameters, resulting in more accurate nowcast predictions. These results support the

notion that incorporating group fixed effects enhances the overall performance of our forecasting model.

Next we address the challenge of missing earnings data, which can complicate the analysis. We examine the performance of parameter imputation methods in computing nowcasts, see, e.g, Brown, Ghysels, and Gredil (2023), even when earnings and/or earnings forecasts are missing for certain observations in the sample. The results obtained through parameter imputation outperform the analyst consensus nowcasts in terms of prediction accuracy.

The paper is organized as follows. Section 2 introduces the models and estimators. A simulation study reporting the finite sample nowcasting performance of our proposed methods appears in Section 3. The results of our empirical application analyzing price-earnings ratios for a panel of individual firms are reported in Section 4. Section 5 concludes. All technical details and detailed data descriptions appear in the Appendix and the Online Appendix.

2 High-dimensional mixed frequency panel data

In this section, we describe the methodological approach of the paper. Motivated by our application, we will refer to the cross-sectional observations as firms, the low-frequency observations as quarterly while the high-frequency observations are daily or monthly. However, the notation presented in this section is generic and can correspond to other entities and frequencies. The objective is to nowcast $\{y_{i,t} : i \in [N], t \in [T]\}$ (where for a positive integer p, we put $[p] = \{1, 2, \ldots, p\}$), in our case a panel of P/E ratios (or its decomposition into returns and analyst forecast errors) for N firms observed at T time periods. The covariates consist of K time-varying predictors measured potentially at higher frequencies

$$\left\{x_{i,t-j/n_k^H,k} : \ i \in [N], t \in [T], j = 0, \ldots, n_k^L n_k^H - 1, k \in [K]\right\},$$

where n_k^H is the number of high-frequency observations for the k^{th} covariate in a low-frequency time period t, and n_k^L is the number of low-frequency time periods used as lags. For instance, $n_k^L = 1$ corresponds in our application to a quarter of high-frequency lags used as covariates and $n_k^H = 3$ corresponds to monthly data with 3 month of data available per quarter. Note that we can think of mixtures of say annual, quarterly, monthly and weekly data, and therefore n_k^H represents different high frequency sampling frequencies and associated lags $n_k^L n_k^H$.

In our empirical analysis we examine three types of regression model specifications: (a) regularized single equation regressions for each individual firm, (b) regularized panel regressions with pooling, and (c) regularized panel regressions with fixed effects. Hence, in (a) we do not explore the panel structure of the data, whereas in (b) and (c) we do. To discuss the model specifications, we focus here on (b) and (c), keeping in mind that the single regression case is a straightforward simplification of the panel regression models.

Consider the mixed frequency panel data regression for $y_{i,t|\tau}$, that is observation i for low-frequency nowcasting y at time t using information up to τ :

$$y_{i,t|\tau} = \alpha_i + \sum_{k=1}^{K} \psi(L^{1/n_k^H}; \beta_k) x_{i,\tau,k} + u_{i,t|\tau},$$

where α_i is the entity-specific intercept (depending on τ but we suppress this detail to simplify notation), and

$$\psi(L^{1/n_k^H}; \beta_k) x_{i,\tau,k} = \frac{1}{k_{max}} \sum_{j=0}^{k_{max}-1} \beta_{j,k} L^{j/n_k^H} x_{i,\tau,k} \tag{1}$$

where k_{max} is the maximum lag length which may depend on the covariate k, and for each high frequency covariate $x_{i,\tau,k}$ we have the most up to date information available at time τ. This may imply that for some high frequency regressors this is stale information as they have not been updated yet, but presumably at least some of the high frequency data are fresh real-time information at the time τ the nowcast is being made. For instance, in our quarterly/monthly application we can have $\tau = (t-1) + 1/3$ in which case we nowcast quarter t with information available at the end of the first month of that quarter. In this example, some high frequency series for the first month may be available while some may not due to say publication lags. Likewise, with $\tau = (t-1) + 2/3$ we can revise the previous nowcast with one extra month of information, which taking into account publication lags may include observations from the first month as the most recent releases. It should parenthetically be noted that for $\tau \leq t-1$, we are dealing with a forecasting situation and therefore our analysis applies to both nowcasting and - ceteris paribus - forecasting.

To reduce the dimensionality of the high-frequency lag polynomial, we follow the MIDAS ML literature, see Babii, Ghysels, and Striaukas (2021, 2022), and estimate a weight function ω parameterized by a relatively small number of coefficients L

$$\psi(L^{1/n_k^H}; \beta_k) x_{i,\tau,k} = \frac{1}{k_{max}} \sum_{j=0}^{k_{max}-1} \omega\left(\frac{j}{n_k^H}; \beta_k\right) x_{i,\tau,k},$$

where the MIDAS weight function is $\omega(s;\beta_k) = \sum_{l=0}^{L-1} \beta_{l,k} w_l(s)$, $(w_l)_{l\geq 0}$ is a collection of L approximating functions, called the *dictionary*, and $\beta_k \in \mathbf{R}^L$ is the unknown parameter. An example of a dictionary used in the MIDAS ML literature is the set of orthogonal Legendre polynomials. To streamline notation it will be convenient to assume, without loss of generality, a common lag length, i.e. $\bar{k}_{max} = k_{max}\ \forall\ k \in [K]$. The linear in parameters dictionaries map the MIDAS regression to a standard linear regression framework. In particular, define $\mathbf{x}_i = (X_{i,1}W, \ldots, X_{i,K}W)$, where for each $k \in [K]$, $X_{i,k} = (x_{i,\tau - j/n_k^H, k}, j = 0, \ldots, \bar{k}_{max} - 1)_{\tau \in [T]}$ is a $T \times \bar{k}_{max}$ matrix of covariates and $\bar{k}_{max} W = (w_l(j/n_k^H; \beta_k)_{0 \leq l \leq L-1, 0 \leq j \leq \bar{k}_{max}}$ is a $\bar{k}_{max} \times L$ matrix corresponding to the dictionary. In addition, let $\mathbf{y}_i = (y_{i,t|\tau}, t, \tau \in [T])^\top$ and $\mathbf{u}_i = (u_{i,t|\tau}, t, \tau \in [T])^\top$. The regression equation after stacking time series observations for each firm $i \in [N]$ is as follows

$$\mathbf{y}_i = \iota \alpha_i + \mathbf{x}_i \beta + \mathbf{u}_i,$$

where $\iota \in \mathbf{R}^T$ is the all-ones vector and $\beta \in \mathbf{R}^{LK}$ is a vector of slope coefficients. Lastly, put $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_N^\top)^\top$, $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top)^\top$, and $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_N^\top)^\top$. Then the regression equation after stacking all cross-sectional observations is

$$\mathbf{y} = B\alpha + \mathbf{X}\beta + \mathbf{u},$$

where $B = I_N \otimes \iota$, I_N is $N \times N$ identity matrix, and $\otimes$ is the Kronecker product. Given that the number of potential predictors K can be large, additional regularization can improve the predictive performance in small samples. To that end, we take advantage of the sg-LASSO regularization, suggested by Babii, Ghysels, and Striaukas (2022).

The fixed effects sg-LASSO estimator $\hat{\rho} = (\hat{\alpha}^\top, \hat{\beta}^\top)^\top$ solves

$$\min_{(a,b) \in \mathbf{R}^{N+p}} \|\mathbf{y} - Ba - \mathbf{X}b\|_{NT}^2 + 2\lambda\Omega(b), \tag{2}$$

where Ω is the sg-LASSO regularizing functional. It is worth stressing that the design matrix $\mathbf{X}$ does not include the intercept and that we do not penalize the fixed effects which are typically not sparse. In addition, $\|.\|_{NT}^2 = |.|^2/(NT)$ is the empirical norm and

$$\Omega(b) = \gamma |b|_1 + (1-\gamma)\|b\|_{2,1},$$

is a regularizing functional. It is a linear combination of the ℓ_1 LASSO and $\ell_{2,1}$ group LASSO norms. Note that for a group structure $\mathcal{G}$ described as a partition of $[p] = \{1, 2, \ldots, p\}$, the group LASSO norm is computed as $\|b\|_{2,1} = \sum_{G \in \mathcal{G}} |b_G|_2$, while $|.|_q$ denotes the usual ℓ_q norm. The group LASSO penalty encourages sparsity between groups whereas the ℓ_1 LASSO norm promotes sparsity within groups and

allows us to learn the shape of the MIDAS weights from the data. The parameter $\gamma \in [0,1]$ determines the relative weights of the ℓ_1 (sparsity) and the $\ell_{2,1}$ (group sparsity) norms, while the amount of regularization is controlled by the regularization parameter $\lambda \geq 0$.

In Section 1, we called our approach structured ML because the group structure allows us to embed the time series structure of the data. More specifically, these structures are represented by groups covering lagged dependent variables and groups of lags for a single (high-frequency) covariate. Throughout the paper, we assume that groups have fixed size, and the group structure is known by the econometrician. Both are reasonable assumptions to make in the context of our empirical application.

For pooled regressions, we assume that all entities share the same intercept parameter $\alpha_1 = \cdots = \alpha_N = \alpha$. The pooled sg-LASSO estimator $\hat{\rho} = (\hat{\alpha}, \hat{\beta}^\top)^\top$ solves

$$\min_{r=(a,b)\in\mathbf{R}^{1+p}} \|\mathbf{y} - a\iota - \mathbf{X}b\|_{NT}^2 + 2\lambda\Omega(r). \tag{3}$$

Pooled regressions are attractive since the effective sample size NT can be huge, yet the heterogeneity of individual time series may be lost. If the underlying series have a substantial heterogeneity over $i \in [N]$, then taking this into account might reduce the projection error and improve the predictive accuracy.

Babii, Ball, Ghysels, and Striaukas (2022) provide the theoretical analysis of predictive performance of regularized panel data regressions with the sg-LASSO regularization, including as special cases (a) standard LASSO, (b) group LASSO regularizations as well as (c) generic high-dimensional panels not involving mixed frequency data. Finally, Babii, Ball, Ghysels, and Striaukas (2022) also develop the debiased inferential methods and Granger causality tests for pooled panel data regressions.

3 Monte Carlo experiments

It is not clear that the aforementioned theory is of practical use in the context of nowcasting using modestly sized samples of data. For this reason, we investigate in this section the finite sample nowcasting performance of the machine learning methods covered so far. We consider the standard (unstructured) elastic net with UMIDAS (called Elnet-U), where UMIDAS refers to unconstrained MIDAS proposed by Foroni, Marcellino, and Schumacher (2015) in a classic non-ML context, and sg-LASSO with MIDAS. Both methods require selecting two tuning parameters λ and γ. In the case of sg-LASSO, γ is the relative weight of LASSO and group LASSO

penalties while in the case of the elastic net γ interpolates between LASSO and ridge. In both cases we report results on a grid $\gamma \in \{0, 0.2, \ldots, 1\}$.

In addition to evaluating the performance over the grid of γ tuning parameter values, we need to select the λ tuning parameter. To do so, we consider several approaches. First, we adapt the K-fold cross-validation to the panel data setting. To that end, we resample the data by blocks respecting the time-series dimension and creating folds based on cross-sectional units instead of the pooled sample. We use the 5-fold cross-validation both in the simulation experiments and the empirical application. We also consider the following three information criteria: BIC, AIC, and corrected AIC (AICc) of Hurvich and Tsai (1989). Assuming that $y_{i,t}|x_{i,t}$ are i.i.d. draws from $N(\alpha_i + x_{i,t}^\top \beta, \sigma^2)$, the log-likelihood of the sample is

$$\mathcal{L}(\alpha, \beta, \sigma^2) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \alpha_i - x_{i,t}^\top \beta)^2.$$

Then, the BIC criterion is

$$\text{BIC} = \frac{\|\mathbf{y} - \hat{\mu} - \mathbf{X}\hat{\beta}\|_{NT}^2}{\hat{\sigma}^2} + \frac{\log(NT)}{NT} \times df,$$

where df denotes the degrees of freedom, $\hat{\sigma}^2$ is a consistent estimator of σ^2, $\hat{\mu} = \hat{\alpha}\iota$ for the pooled regression, and $\hat{\mu} = B\hat{\alpha}$ for fixed effects regression. The degrees of freedom are estimated as $\widehat{df} = |\hat{\beta}|_0 + 1$ for the pooled regression and $\widehat{df} = |\hat{\beta}|_0 + N$ for the fixed effects regression, where $|.|_0$ is the ℓ_0-norm defined as a number of non-zero coefficients; see Zou, Hastie, and Tibshirani (2007) for more details. The AIC is computed as

$$\text{AIC} = \frac{\|\mathbf{y} - \hat{\mu} - \mathbf{X}\hat{\beta}\|_{NT}^2}{\hat{\sigma}^2} + \frac{2}{NT} \times \widehat{df},$$

and the corrected Akaike information criteria is

$$\text{AICc} = \frac{\|\mathbf{y} - \hat{\mu} - \mathbf{X}\hat{\beta}\|_{NT}^2}{\hat{\sigma}^2} + \frac{2\widehat{df}}{NT - \widehat{df} - 1}.$$

The AICc is typically a better choice when p is large relative to the sample size. We report the results for each of the tuning parameter selection criteria for λ, along the grid choice for γ.

3.1 Simulation Design

To assess the predictive performance of pooled panel data models, we simulate the data from the following DGP with a quarterly/monthly frequency mix in mind and $\bar{k}_{max} = k_{max}$ with $n_k^H = n^H \; \forall \; k$:

$$y_{i,t|\tau} = \alpha + \sum_{k=1}^{K} \bar{k}_{max}^{-1} \sum_{j=0}^{\bar{k}_{max}-1} \omega(j/n^H; \beta_k) x_{i,\tau-j/n^H,k} + u_{i,t|\tau},$$

where $i \in [N]$, $t \in [T]$, α is the common intercept, $\bar{k}_{max}^{-1} \sum_{j=0}^{\bar{k}_{max}-1} \omega(j/n_k; \beta_k)$ the weight function for k-th high-frequency covariate and the error term is either $u_{i,t|\tau} \sim_{i.i.d.} N(0,1)$ or $u_{i,t|\tau} \sim_{i.i.d.}$ student-$t(5)$.

We are interested in a quarterly/monthly data mix, and use four quarters of data for the high-frequency regressors which covers 12 high-frequency lags for each regressor. In terms of information sets we start with $\tau = t-1$, which corresponds to a prediction setting and then have $\tau = t-1+1/3$, i.e. nowcasting with one month's worth of information. We set the number of relevant high-frequency regressors $K = 6$. The high-frequency regressors are generated as K i.i.d. realizations of the univariate autoregressive (AR) process $x_h = \rho x_{h-1} + \varepsilon_h$, where $\rho = 0.6$ and either $\varepsilon_h \sim_{i.i.d.} N(0,1)$ or $\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$, where h denotes the high-frequency sampling. We rely on a commonly used weighting scheme in the MIDAS literature, namely $\omega(s; \beta_k)$ for $k = 1, 2, \ldots, 6$ are determined by beta densities respectively equal to Beta$(1,3)$ for $k = 1, 4$, Beta$(2,3)$ for $k = 2, 5$, and Beta$(2,2)$ for $k = 3, 6$; see Ghysels, Sinko, and Valkanov (2007) or Ghysels and Qian (2019), for further details. The MIDAS regressions are estimated using Legendre polynomials of degree $L = 3$.

We consider DGPs featuring pooled panels and fixed effects. For the pooled panel regression DGPs we simulate the intercepts as $\alpha \sim$ Uniform$(-4, 4)$. For the fixed effects models the individual fixed effects are simulated as $\alpha_i \sim_{\text{i.i.d}}$ Uniform$(-4, 4)$ and are kept fixed throughout the experiment.

For $\tau = t-1$, the *Baseline scenario*, in the estimation procedure we add 24 noisy covariates which are generated in the same way as the relevant covariates, use 4 low-frequency lags and the error terms $u_{i,t|\tau}$ and ε_h are Gaussian. In the student-$t(5)$ scenario we replace the Gaussian error terms with a student-$t(5)$ distribution while in the *large dimensional* scenario we add 94 noisy covariates. For each scenario, we simulate $N = 25$ i.i.d. time series of length $T = 50$; next we increase the cross-sectional dimension to $N = 75$ and time series to $T = 100$.

Finally, for $\tau = t-1+1/3$ the thought experiment in the simulation design is one where the first high-frequency observations during low frequency t are available.

The nowcaster of course does not know which of the covariates are relevant nor does she know the parameters of the prediction rule. We will call this scheme "one-step ahead" nowcasts.

3.2 Simulation results

Tables 1 and 2 cover the average mean squared forecast errors (MSFE) for one-step ahead nowcasts for the three simulation scenarios. We report results for sg-LASSO with MIDAS weights (left block) and elastic net with UMIDAS (right block) using both pooled panel models (Table 1) and fixed effects ones (Table 2). We report results for the best choice of the γ tuning parameter.[1]

Firstly, structured sg-LASSO-MIDAS consistently outperforms unstructured Elnet-U for all DGPs and in both pooled and fixed effects cases. The most significant discrepancy between the two methods is observed in situations with small N and small T, specifically when N = 25 and T = 50. As either N or T increases, this gap gradually diminishes. When comparing the results of pooled and fixed effects, it becomes evident that the difference between the two approaches — structured sg-LASSO-MIDAS versus Elnet UMIDAS — widens further in the case of fixed effects with student-t(5) data. This indicates that our structured approach yields higher quality estimates for the fixed effects and thus more accurate nowcasts.

In the case of sg-LASSO-MIDAS, the best performance is achieved for $\gamma \notin \{0, 1\}$ for both pooled panel data and fixed effects cases, while $\gamma = 0$, i.e. ridge regression, seems to be dominated by estimators that $\gamma \notin \{0, 1\}$ in both pooled and fixed effects cases. For the student-$t(5)$ and large dimensional DGP, we observe a decrease in the performance for all methods. However, the decrease in the performance is larger for the student-$t(5)$ DGP, revealing that heavy-tailed data have — as expected — a stronger impact on the performance of the estimators.

For the pooled panel data case, increasing N from 25 to 75 seems to have a larger positive impact on the performance than an increase in the time-series dimension from $T = 50$ to $T = 100$. The difference appears to be larger for student-$t(5)$ and large dimensional DGPs and/or for the elastic net case. Turning to the fixed effects results, the differences seem to be even sharper, in particular for student-$t(5)$ and large dimensional DGPs.

When comparing the results across the different model selection methods, i.e., cross-validation and the three information criteria, we find that almost always cross-

[1] Results for the grid of $\gamma \in \{0.0, 0.2, \ldots, 1.0\}$ are reported in the Online Appendix Tables OA.1-OA.3.

	sg-LASSO			Elnet-U		
N/T =	25/50	75/50	25/100	25/50	75/50	25/100
			Panel A. Baseline			
CV	1.191	1.157	1.168	1.213	1.158	1.172
BIC	1.270	1.175	1.202	1.384	1.211	1.247
AIC	1.234	1.160	1.187	1.273	1.172	1.213
AICc	1.237	1.161	1.188	1.279	1.172	1.217
			Panel B. Student-$t(5)$			
CV	1.280	1.245	1.248	1.299	1.243	1.256
BIC	1.389	1.274	1.293	1.570	1.317	1.367
AIC	1.345	1.259	1.272	1.411	1.283	1.298
AICc	1.344	1.259	1.273	1.412	1.283	1.300
			Panel C. Large-dimensional			
CV	1.204	1.160	1.185	1.255	1.165	1.188
BIC	1.273	1.175	1.214	1.409	1.208	1.289
AIC	1.259	1.166	1.191	1.350	1.198	1.232
AICc	1.260	1.167	1.192	1.353	1.200	1.232

Table 1: The table reports the MSFE for nowcasting accuracy for the pooled estimator for the Baseline (Panel A), student-$t(5)$ (Panel B), and large-dimensional (Panel C) DGPs for the sg-LASSO-MIDAS (rows sg-LASSO) and elastic net UMIDAS (rows Elnet-U). We vary the cross-sectional dimension $N \in \{25, 75\}$ and time series dimension $T \in \{50, 100\}$. We report results for 5-fold cross-validation, BIC, AIC, AICc information criteria λ tuning parameter calculation methods and for the best choice of γ tuning parameter.

	sg-LASSO			Elnet-U		
N/T =	25/50	75/50	25/100	25/50	75/50	25/100
			Panel A. Baseline			
CV	1.198	1.170	1.164	1.245	1.183	1.184
BIC	1.304	1.202	1.213	1.537	1.259	1.313
AIC	1.282	1.192	1.196	1.380	1.222	1.237
AICc	1.284	1.193	1.196	1.284	1.193	1.196
			Panel B. Student-$t(5)$			
CV	1.278	1.256	1.248	1.329	1.270	1.271
BIC	1.437	1.306	1.310	1.694	1.367	1.404
AIC	1.389	1.292	1.294	1.478	1.316	1.342
AICc	1.393	1.293	1.295	1.495	1.316	1.348
			Panel C. Large-dimensional			
CV	1.214	1.170	1.172	1.282	1.197	1.193
BIC	1.344	1.213	1.229	1.662	1.298	1.342
AIC	1.300	1.243	1.202	1.404	1.384	1.235
AICc	1.301	1.205	1.204	1.405	1.247	1.238

Table 2: The table reports the MSFE for nowcasting accuracy for the fixed effects estimator for the Baseline (Panel A), student-$t(5)$ (Panel B), and large-dimensional (Panel C) DGPs for the sg-LASSO-MIDAS (rows sg-LASSO) and elastic net UMIDAS (rows Elnet-U). We vary the cross-sectional dimension $N \in \{25, 75\}$ and time series dimension $T \in \{50, 100\}$. We report results for 5-fold cross-validation, BIC, AIC, AICc information criteria λ tuning parameter calculation methods and for the best choice of γ tuning parameter.

validation leads to smaller prediction errors in both pooled and fixed effects panel data cases. Notably, the gains appear to be larger for the large N and T values. Comparing BIC, AIC, and AICc information criteria, the results appear to be similar for AIC and AICc across DGPs and different sample sizes, while the BIC performance is slightly worse than AIC and AICc.

4 Nowcasting price-earnings ratios

Ball and Ghysels (2018), Carabias (2018) and Babii, Ball, Ghysels, and Striaukas (2022) documented that analysts make systematic and predictable errors in their P/E forecasts. We therefore consider nowcasting the P/E ratios using a set of predictors that are sampled at mixed frequencies for a large cross-section of firms.

A natural question one may ask: should we nowcast P/E ratio directly or it's components. We, therefore, decompose the (log of) the P/E ratio for firm i as follows:

$$\begin{aligned} pe_{i,t+1} \equiv \log(P_{i,t+1}/E_{i,t+1}) &= \log((P_{i,t+1}/P_{i,t})/(E_{i,t+1}/P_{i,t})) \\ &= r_{i,t+1} - \log((E_{i,t+1}/E^a_{i,t+1|t})/(P_{i,t}/E^a_{i,t+1|t})) \\ &= r_{i,t+1} - e^a_{i,t+1|t} + \log(P_{i,t}/E^a_{i,t+1|t}) \end{aligned} \tag{4}$$

where $r_{i,t+1}$ is the log return from $t+1$ to t for firm i, $E^a_{i,t+1|t}$ the analyst's prediction at time t pertaining to $t+1$ earnings, and $e^a_{i,t+1|t} \equiv \log(E_{i,t+1}) - \log(E^a_{i,t+1|t})$ is the log earnings forecast error of analysts pertaining to their end of period t prediction for $t+1$. Finally, $\log(P_{i,t}/E^a_{i,t+1|t})$ is perfectly known at time t. The above defines an additive decomposition of the log P/E ratio into the return for firm i and the analyst prediction error. Therefore, nowcasting the log P/E ratio could also be achieved via nowcasting its two components. The decomposition corresponds to the distinction between analyst assessments of firm i's earnings and market/investor assessments of the firm.

There is a considerable literature on using machine learning to predict returns, see e.g. Rapach, Strauss, and Zhou (2010), Kim and Swanson (2014), Gu, Kelly, and Xiu (2020), D'Hondt, De Winne, Ghysels, and Raymond (2020), among others. Here we are dealing with a slightly modified setting where we are nowcasting quarterly returns with information during quarter $t+1$. Nevertheless, prediction and nowcasting are closely related. The second component, $e^a_{i,t+1|t}$ has been explored by Babii, Ball, Ghysels, and Striaukas (2022), who revisit a topic raised by Ball and Ghysels (2018) and Carabias (2018), and confirmed in a rich data setting that analysts tend to focus

on their firm/industry when making earnings predictions while not fully taking into account the impact of macroeconomic events. Put differently, one can forecast and nowcast analyst prediction errors.

It should also parenthetically be noted that equation (4) can be rewritten as a decomposition of returns, namely:

$$r_{i,t+1} = pe_{i,t+1} + e^a_{i,t+1|t} + \log(P_{i,t}/E^a_{i,t+1|t}) \tag{5}$$

which can be viewed as an alternative decomposition of returns compared to Ferreira and Santa-Clara (2011). They propose forecasting separately the three components of stock market returns: (a) the dividend price ratio, (b) earnings growth, and (c) price-to-earnings ratio growth. Ferreira and Santa-Clara (2011) argue that predicting the separate components yields better return predictions compared to the usual models producing direct forecasts of the latter. They estimate the expected earnings growth using a 20-year moving average of the growth in earnings per share. The expected dividend price ratio is estimated by the current dividend price ratio. This implicitly assumes that the dividend price ratio follows a random walk. While our application is different in many regards, the arguments being considered are similar. It is worth reminding ourselves that if the nowcast $\widehat{pe}_{i,t+1}$ is constructed from individual component nowcasts, then

$$\text{MSE}(\widehat{pe}_{i,t+1}) = \text{MSE}(\widehat{r}_{i,t+1}) + \text{MSE}(\hat{e}^a_{i,t+1|t}) - 2\mathbb{E}\left[(r_{i,t+1} - \hat{r}_{i,t+1})(e^a_{i,t+1|t} - \hat{e}^a_{i,t+1|t})\right] \tag{6}$$

Hence, depending on the co-movements between returns for firm i, $r_{i,t+1}$ and analyst earning prediction errors $e^a_{i,t+1|t}$, we are better off to directly predict $pe_{i,t+1}$ or its components. If the latter are positively correlated, then we are better off direct forecasting is preferred.

Given the aforementioned decomposition, we are interested in the following LHS variables: $pe_{i,t+1}$, $r_{i,t+1}$ and $e^a_{i,t+1|t}$. First, we estimate the individual sg-LASSO MIDAS regressions for each firm $i = 1, \ldots, N$, namely:

$$\mathbf{y}_i = \iota\alpha_i + \mathbf{x}_i\beta_i + \mathbf{u}_i,$$

where the firm-specific predictions are computed as $\hat{y}_{i,t+1} = \hat{\alpha}_i + x^\top_{i,t+1}\hat{\beta}_i$. As noted in Section 2, $\mathbf{x}_i$ contains lags of the low-frequency target variable and high-frequency covariates to which we apply Legendre polynomials of degree $L = 3$.

Next, we estimate the following pooled and fixed effects sg-LASSO MIDAS panel data models

$$\begin{aligned} \mathbf{y} &= \alpha\iota + \mathbf{X}\beta + \mathbf{u} \quad \text{Pooled} \\ \mathbf{y} &= B\alpha + \mathbf{X}\beta + \mathbf{u} \quad \text{Fixed Effects} \end{aligned}$$

and compute predictions as

$$\begin{aligned} \hat{y}_{i,t+1} &= \hat{\alpha} + x_{i,t+1}^{\top}\hat{\beta} \quad \text{Pooled} \\ \hat{y}_{i,t+1} &= \hat{\alpha}_i + x_{i,t+1}^{\top}\hat{\beta} \quad \text{Fixed Effects.} \end{aligned}$$

Once we compute the forecast for the log of P/E ratio ($pe_{i,t+1}$), log returns ($r_{i,t+1}$) and log earnings forecast error ($e^a_{i,t+1|t}$), we compute the final prediction accuracy metrics by either taking directly log P/E nowcast or the sum of its components, i.e., $\hat{S} = \hat{r}_{i,t+1} - \hat{e}^a_{i,t+1|t} + \log(P_{i,t}/E^a_{i,t+1|t})$.

We benchmark firm-specific and panel data regression-based nowcasts against two simple alternatives. First, we compute forecasts for the RW model as

$$\hat{y}_{i,t+1|t} = y_{i,t}.$$

Second, we consider predictions of P/E implied by analysts' earnings nowcasts using the information up to time $t+1$, i.e.

$$\hat{y}_{i,t+1|t} = \bar{y}^a_{i,t+1|t},$$

where the predicted/nowcasted log of P/E ratio is based on consensus earnings forecasts pertaining to the end of the $t+1$ quarter using the stock price at the end of quarter t. To measure the forecasting performance, we compute the mean squared forecast errors (MSE) for each method. Let $\bar{\mathbf{y}}_i = (y_{i,T_{is}+1}, \ldots, y_{i,T_{os}})^{\top}$ represent the out-of-sample realized P/E ratio values, where T_{is} and T_{os} denote the last in-sample observation for the first prediction and the last out-of-sample observation respectively, and let $\hat{\mathbf{y}}_i = (\hat{y}_{i,t_{is}+1}, \ldots, \hat{y}_{i,t_{os}})$ collect the out-of-sample forecasts. Then, the mean squared forecast errors are computed as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{T - T_{is} + 1}(\bar{\mathbf{y}}_i - \hat{\mathbf{y}}_i)^{\top}(\bar{\mathbf{y}}_i - \hat{\mathbf{y}}_i).$$

We look at 210 US firms and use 24 predictors, including traditional macro and financial series as well as non-traditional series from textual analysis of financial news. We apply (a) single regression individual firm high-dimensional regressions, (b) pooled and (c) individual fixed effects sg-LASSO MIDAS panel data models and report results for several choices of the tuning parameters. We compare these three type of models with several benchmarks, which include a random walk (RW) model and analysts' consensus forecasts. The remainder of the section is structured as follows. We start with a short review of the data followed by a summary of the empirical results.

4.1 Data description

The full sample consists of observations between the 1^{st} of January, 2000 and the 30^{th} of June, 2017. Due to the lagged dependent variables in the models, our effective sample starts in the third fiscal quarter of 2000. We use the first 25 observations for the initial sample, and use the remaining 42 observations for evaluating the out-of-sample forecasts, which we obtain by using an expanding window forecasting scheme. We collect data from CRSP and I/B/E/S to compute the quarterly P/E ratios and firm-specific financial covariates; RavenPack is used to compute daily firm-level textual-analysis-based data; real-time monthly macroeconomic series are from the FRED-MD dataset, see McCracken and Ng (2016) for more details; FRED is used to compute daily financial markets data and, lastly, monthly news attention series extracted from the *Wall Street Journal* articles are retrieved from Bybee, Kelly, Manela, and Xiu (2021).[2] Online Appendix Section OA.2 provides a detailed description of the data sources.[3]

Our target variable is the P/E ratio for each firm. To compute it, we use CRSP stock price data and I/B/E/S earnings data. Earnings data are subject to release delays of 1 to 2 months depending on the firm and quarter. Therefore, to reflect the real-time information flow, we compute the target variable using stock prices that are available in real-time. We also take into account that different firms have different fiscal quarters, which also affects the real-time information flow.

For example, suppose for a particular firm the fiscal quarters are at the end of the third month in a quarter, i.e. end of March, June, September, and December. The consensus forecast of the P/E ratio is computed using the same end-of-quarter price data which is divided by the earnings consensus forecast value. The consensus is computed by taking all individual prediction values up to the end of the quarter and aggregating those values by taking either the mean or the median. To compute the target variable, we adjust for publication lags and use prices of the publication date instead of the end of fiscal quarter prices. More precisely, suppose we predict the P/E ratio for the first quarter. As noted earlier, earnings are typically published with 1 to 2 months delay; say for a particular firm the data is published on the $25th$ of April. In this case, we record the stock price for the firm on $25th$ of April, and divide it by the earnings announced on that date.

[2]The dataset is publicly available at http://www.structureofnews.com/.

[3]In particular, firm-level variables, including P/E ratios, are described in Online Appendix Table OA.4, and the other predictor variables in Online Appendix Table OA.5. The list of all firms we consider in our analysis appears in Online Appendix Table OA.6.

4.2 Models and main results

To simplify the exposition, we denote y as one of the three target variables we consider. The main findings from our analysis are presented in Table 3. Column $\hat{pe}_{i,t+1}$ reports results for directly nowcasting the log P/E ratio, column $\hat{S}$ reports the results of nowcasting and summing up the components, column $r_{i,t+1}$ reports results for the log return component and column $\hat{e}^a_{i,t+1|t}$ reports results for the log earnings forecast error of analysts component. Row *RW* reports results for the random walk, while row *Consensus* for the median consensus nowcast. Panels *Individual*, *Pooled* and *Fixed effects* report results for different panel data models relative to the consensus MSE (columns $\hat{pe}_{i,t+1}$ and $\hat{S}$) and for the components (columns $r_{i,t+1}$ and $\hat{e}^a_{i,t+1|t}$) we report ratios relative to the RW MSE since there are obviously no concensus series notably for the analyst forecast errors.

Nowcasting Performance

In light of the simulation evidence, we report the empirical results using cross-validation in Table 3 and provide the full set of results in Online Appendix Table OA.7. The entries in the top panel of Table 3 reveal that predictions based on analyst consensus exhibit significantly higher mean squared forecast errors (MSEs) compared to model-based predictions since all the ratios with respect to the concensus are less than one (see first two columns). These model-based predictions involve either direct log P/E ratio nowcasts (first column) or their individual components (second column). Since the MSE for RW and concensus are quite similar, this also implies that RW predictions are outperformed by the model-based ones. The substantial improvement in the accuracy of model-based predictions compared to analyst-based predictions underscores the value of employing machine learning techniques for nowcasting log P/E ratios. Across various machine learning methods, including single-firm and panel data regressions, we consistently observe enhanced performance. When comparing the first and second columns, which correspond to direct log P/E ratio nowcasts versus those based on its components, we observe a substantial enhancement in prediction accuracy when using the individual components. This improvement is consistently evident across individual, pooled, and fixed effects regression models. To shed light on these findings, we computed the pooled correlation between returns and earnings for the entire sample, i.e. $\text{Corr}(r_{i,t+1}, e^a_{i,t+1|t})$ = -0.206. The correlation indicates a (weak) negative relationship between returns and earnings. Consequently, the prediction errors of each component tend to offset each other, resulting in more accurate aggregated nowcasts (recall equation (6)). The last two columns of Table 3 present the prediction results for these components.

	$\hat{pe}_{i,t+1}$	$\hat{S}$	$\hat{r}_{i,t+1}$	$\hat{e}^a_{i,t+1\|t}$
			All firms	
RW	1.355		0.054	0.194
Consensus	1.305			
			Individual	
	0.905	0.890	1.088	0.848
DM p-val RW	0.117	0.115	0.181	0.090
DM p-val Cons.	0.156	0.131		
			Pooled	
	0.894	0.790	0.964	0.799
DM p-val RW	0.060	0.023	0.128	0.021
DM p-val Cons.	0.075	0.053		
			Fixed effects	
	0.814	0.793	0.971	0.803
DM p-val RW	0.051	0.033	0.164	0.032
DM p-val Cons.	0.078	0.063		
	With single CCI outlier removed (see Figure 2)			
RW	1.333		0.053	0.173
Consensus	1.275			
			Individual	
	0.978	0.790	1.001	0.812
DM p-val RW	0.585	0.027	0.912	0.081
DM p-val Cons.	0.606	0.034		
			Pooled	
	0.777	0.768	0.943	0.788
DM p-val RW	0.025	0.004	0.103	0.018
DM p-val Cons.	0.029	0.006		
			Fixed effects	
	0.782	0.767	0.954	0.783
DM p-val RW	0.028	0.004	0.119	0.021
DM p-val Cons.	0.030	0.006		

Table 3: Column $\hat{pe}_{i,t+1}$ reports results for directly nowcasting the log P/E ratio, $\hat{S}$ for nowcasting and summing up the components, $r_{i,t+1}$ for the log return and $\hat{e}^a_{i,t+1|t}$ for the log earnings forecast error of analysts. *RW* is for the random walk, while *Consensus* is the median consensus nowcast. Panels *Individual*, *Pooled* and *Fixed effects* report results for models relative to the consensus MSE ($\hat{pe}_{i,t+1}$ and $\hat{S}$) and for the components ($r_{i,t+1}$ and $\hat{e}^a_{i,t+1|t}$) relative to the RW MSE. DM is the Diebold and Mariano (1995) test statistic p-values using one-sided critical values.

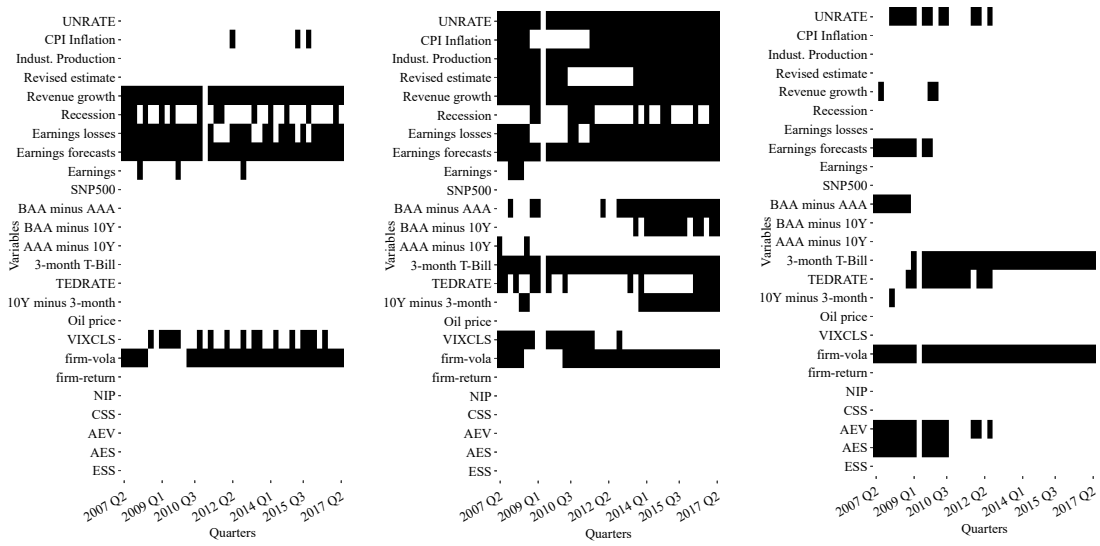
We observe that analyst earnings prediction errors appear to be more predictable than those of log returns. We also report Diebold and Mariano (1995) test statistic p-values comparing each model against the RW and consensus benchmarks, pooling all the nowcasting errors across firms. Using one-sided test critical values we observe that our models outperform both the RW and consensus benchmarks, particularly when we use the component approach. While we cannot compare the $\hat{p}e_{i,t+1}$ component with the consensus, judging by the RW benchmark it is clear that the second component is the most important in terms of nowcasting gains. When we use individual MIDAS regressions the evidence is less compelling, underscoring the importance of using panel data models.[4]

Sparsity Patterns

Figure 1 illustrates the sparsity patterns of selected covariates for the most effective methods in predicting either log P/E ratios (Panel a) or their components (Panels b and c). It is worth noting that the sparsity patterns differ significantly across the three panels. For instance, firm volatility is often chosen as a relevant covariate across all targets, albeit not consistently throughout the entire out-of-sample period. In the case of log P/E ratios, news series related to earnings are frequently selected, along with firm and market volatility series. Conversely, for log returns, a denser pattern of covariate selection is observed, distinct from the other two cases. Interestingly, none of the news-based firm series are chosen for this target. Regarding log analyst earnings forecast errors, macroeconomic series such as the unemployment rate, short-term rates, and TED rate are frequently selected. Moreover, unlike log P/E ratios and returns, news-based firm series occasionally appear in the selected covariates for this target. The fact that macroeconomic series are drivers for nowcasting the $e^a_{i,t+1|t}$ component is a confirmation of the findings reported in Ball and Ghysels (2018), Carabias (2018) and Babii, Ball, Ghysels, and Striaukas (2022).

Figure 2 depicts the histogram of mean squared errors (MSEs) across firms. Notably, a substantial proportion of firms (approximately 60%) exhibit low MSE values, indicating a high level of prediction accuracy. However, there are a few firms for which the MSEs are relatively larger, suggesting lower prediction performance for

[4]We also experimented with the forecast combination of MIDAS regressions used by Ball and Ghysels (2018) and found them to be inferior to the individual MIDAS ML regressions as well as the panel data models. We therefore refrain from reporting the details here. In addition, we computed prediction results using LASSO estimator with covariates which are averages within the quarter of the high frequency regressors, which we refer to as LASSO-AGG. In all but one of the cases reported in Table 3, the ratio for LASSO-AGG is larger than 1 compared to the corresponding sg-LASSO with MIDAS, indicating that the latter improves out-of-sample results over LASSO-AGG.



(a) Best model for $\hat{pe}_{i,t+1}$. (b) Best model for $\hat{r}_{i,t+1}$. (c) Best model for $\hat{e}^{a}_{i,t+1|t}$.

Figure 1: Sparsity patterns.

these specific cases. The largest MSE is for Crown castle international corporation (CCI) which appears as a strong outlier.

Removing the single outlier firm has a dramatic impact on the nowcasting performance evaluation as shown in the lower panel of Table 3. We now have very strong evidence that the panel regression models dominate analyst predictions. Again the component nowcasts are the best, but even the individual regression models do significantly better when the component specification is used.

Daily Updates of Nowcasts

Our framework allows us to go beyond providing quarterly nowcasts and generate daily updates of earnings series. Leveraging the daily influx of information throughout the quarter, we continuously re-estimate our models and produce nowcast updates as soon as new data becomes available. In Figure 3, we present the distribution of Mean Squared Errors (MSEs) across firms for five distinct nowcast horizons: 20-day, 15-day, 10-day, and 5-day ahead, as well as the end of the quarter. We report the best model based on Table 3. Notably, as the horizons become shorter, both the median and upper quartile of MSEs decrease. Therefore, updating nowcasts with daily information appears to significantly enhance the prediction performance of log earnings ratios. The largest errors persist for the same firm, CCI.

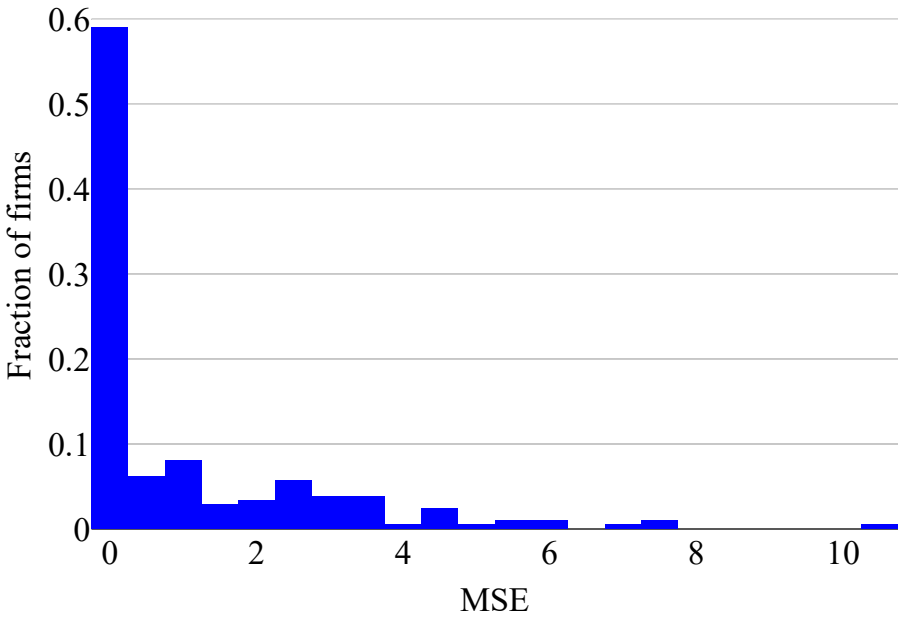


Figure 2: Histogram of mean squared errors.

Grouped Fixed Effects based on
Fama-French Industry Classification

The sg-LASSO estimator we employ in our study is well-suited for incorporating grouped fixed effects. This approach involves grouping firm-specific intercepts based on either statistical procedures or economic reasoning, as outlined in Bonhomme and Manresa (2015). In our analysis, we utilize the Fama French industry classification to form 10 distinct groups for grouping fixed effects. Rather than assuming a common fixed effect for all firms within a group, we apply a group penalty to the fixed effects of firms belonging to the same industry. This allows us to capture industry-specific heterogeneity while avoiding overfitting.

We present the findings in Table 4, which highlight several key observations. Similar to previous analyses, our results suggest that predicting individual components of the log price-earnings ratio leads to more accurate aggregate nowcasts compared to a direct nowcast approach. Furthermore, we observe that the use of group fixed effects improves the accuracy of our nowcasts when forecasting individual components. This can be seen in column 2 of both Tables 3 and 4. Comparatively, when considering the best tuning parameter choice, grouped fixed effects outperform other

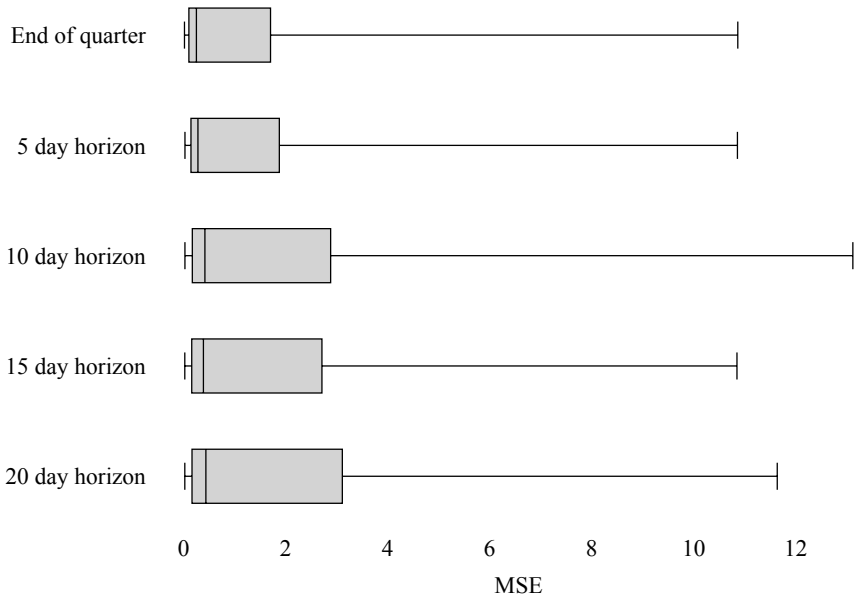


Figure 3: Distribution of MSEs of the best performing model in Table 3. Models are re-estimated for each horizon. The best model based on Table 3 is reported.

panel models, including the pooled panel model. Therefore, our findings suggest that grouped fixed effects strike a better balance between capturing heterogeneity and pooled parameters, resulting in more accurate nowcast predictions. These results support the notion that incorporating group fixed effects enhances the overall performance of our forecasting model.

In Figure 4, we present the distribution of (MSEs) across firms for five industries, based on the best model specification from Table 4. The industries we focus on are the ones with the highest number of firms in our sample. The results reveal variations in performance among different industries. Specifically, the firms categorized as *Consumer Durables* exhibit the lowest accuracy in terms of the median MSE, although the quartiles are comparatively lower compared to the other industries. On the other hand, the nowcasts for firms in the *Consumer Nondurables* and *Others* categories demonstrate the highest accuracy at the median. However, it is important to note that the largest errors occur within the firms classified as *Others*.

Nowcasting with Missing Data — Parameter Imputation Method

Next we address the challenge of missing earnings data, which can complicate

	$\hat{pe}_{i,t+1}$	$\hat{S}$
	Group fixed effects	
CV	0.862	0.789
BIC	0.834	0.789
AIC	0.842	0.791
AICc	0.842	0.790

Table 4: Nowcasting results. Column $\hat{pe}_{i,t+1}$ reports results for directly nowcasting the log P/E ratio and the column $\hat{S}$ reports the results of nowcasting and summing up the components. Results are reported relative to the *Consensus* nowcasts that appear in Table 3.

the analysis. We examine the performance of parameter imputation methods in computing nowcasts, see, e.g, Brown, Ghysels, and Gredil (2023), even when earnings and/or earnings forecasts are missing for certain observations in the sample. We identify a subset of 117 firms for which at least one earnings observation is available in our out-of-sample period, and for which we have matched daily news data. To handle missing data, we match these firms with missing observations to firms in our main sample using the Fama French industry classification. We then utilize the parameter estimates obtained from the best group fixed effects model, as shown in Table 4, to compute the nowcasts of log earnings ratios, either directly or based on its components. The results of this analysis appear in Table 5.

Firstly, the results obtained through parameter imputation support the conclusion that nowcasting the components of the log earnings ratio yields higher quality predictions. This indicates that incorporating the individual components of the ratio improves the accuracy of the nowcasts. Secondly, the panel models with the parameter imputation method outperform the analyst consensus nowcasts in terms of prediction accuracy. This suggests that employing machine learning panel data models along with parameter imputation could be a straightforward yet effective approach in situations where earnings data is not available. Overall, these findings highlight the potential benefits of leveraging machine learning techniques and imputation methods for improving nowcasting accuracy, particularly in cases where earnings data may be missing.

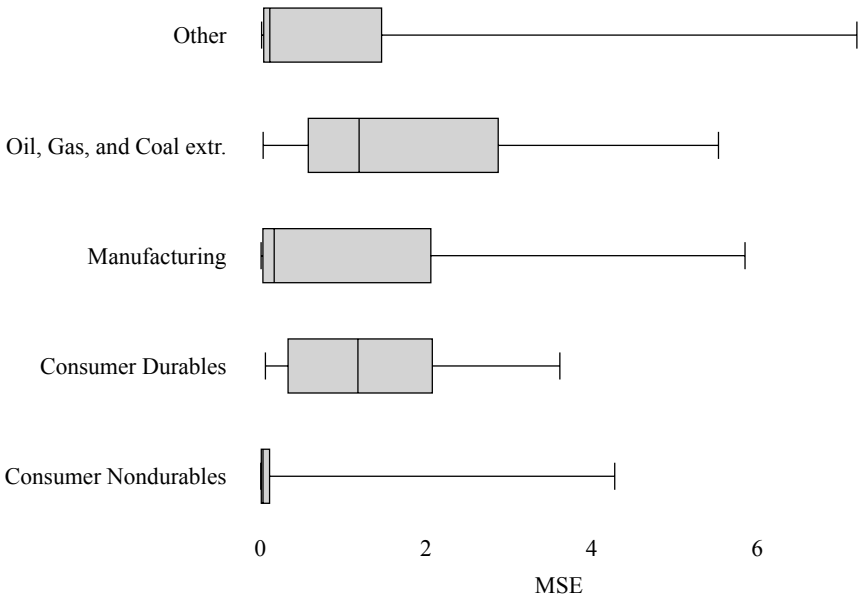


Figure 4: Distribution of MSEs for five industries based on Fama French classification. The reported resutls are based on the best model specification from Table 4.

5 Conclusions

This paper uses a new class of high-dimensional panel data nowcasting models with dictionaries and sg-LASSO regularization which is an attractive choice for the predictive panel data regressions, where the low- and/or the high-frequency lags define a clear group structure. Our empirical results showcase the advantages of using regularized panel data regressions for nowcasting corporate earnings either directly or using a decomposition which separates stock market return predictions and analyst assessments of a firm's performance. While nowcasting earnings is a leading example of applying panel data MIDAS machine learning regressions, one can think of many other applications of interest in finance. Beyond earnings, analysts are also interested in sales, dividends, etc. Our analysis can also be useful for other areas of interest, such as regional and international panel data settings.

	$\hat{pe}_{i,t+1}$	$\hat{S}$
Consensus	1.605	
CV	0.883	0.753
BIC	0.877	0.756
AIC	0.883	0.754
AICc	0.883	0.753

Table 5: Nowcasting results — parameter imputation method. Column $\hat{pe}_{i,t+1}$ reports results for directly nowcasting the log P/E ratio and the column $\hat{S}$ reports the results of nowcasting and summing up the components. Row *Consensus* for the median consensus nowcast. Panels *Individual*, *Pooled* and *Fixed effects* report results for different panel data models relative to the consensus MSE.

References

Babii, A., R. T. Ball, E. Ghysels, and J. Striaukas (2022): "Machine Learning Panel Data Regressions with Heavy-tailed Dependent Data: Theory and Application," *Journal of Econometrics*, (forthcoming).

Babii, A., E. Ghysels, and J. Striaukas (2021): "High-dimensional Granger causality tests with an application to VIX and news," *Journal of Financial Econometrics*, (forthcoming).

——— (2022): "Machine learning time series regressions with an application to nowcasting," *Journal of Business and Economic Statistics*, 40, 1094–1106.

Ball, R. T., and E. Ghysels (2018): "Automated earnings forecasts: beat analysts or combine and conquer?," *Management Science*, 64, 4936–4952.

Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013): "Nowcasting and the real-time data flow," in *Handbook of Economic Forecasting, Volume 2 Part A*, ed. by G. Elliott, and A. Timmermann, pp. 195–237. Elsevier.

Bonhomme, S., and E. Manresa (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83(3), 1147–1184.

Brown, G. W., E. Ghysels, and O. R. Gredil (2023): "Nowcasting net asset values: The case of private equity," *The Review of Financial Studies*, 36(3), 945–986.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2021): "Business News and Business Cycles," Available at SSRN 3446225.

Carabias, J. M. (2018): "The real-time information content of macroeconomic news: implications for firm-level earnings expectations," *Review of Accounting Studies*, 23(1), 136–166.

Diebold, F. X., and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13(3), 253–263.

D'Hondt, C., R. De Winne, E. Ghysels, and S. Raymond (2020): "Artificial intelligence alter egos: Who might benefit from robo-investing?," *Journal of Empirical Finance*, 59, 278–299.

Ferreira, M. A., and P. Santa-Clara (2011): "Forecasting stock market returns: The sum of the parts is more than the whole," *Journal of Financial Economics*, 100(3), 514–537.

Foroni, C., M. Marcellino, and C. Schumacher (2015): "Unrestricted mixed data sampling (U-MIDAS): MIDAS regressions with unrestricted lag polynomials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1), 57–82.

Fosten, J., and R. Greenaway-McGrevy (2019): "Panel data nowcasting," *Available at SSRN 3435691*.

Ghysels, E., and H. Qian (2019): "Estimating MIDAS regressions via OLS with polynomial parameter profiling," *Econometrics and Statistics*, 9, 1–16.

Ghysels, E., A. Sinko, and R. Valkanov (2007): "MIDAS regressions: Further results and new directions," *Econometric Reviews*, 26(1), 53–90.

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

Hurvich, C. M., and C.-L. Tsai (1989): "Regression and time series model selection in small samples," *Biometrika*, 76(2), 297–307.

Khalaf, L., M. Kichian, C. J. Saunders, and M. Voia (2021): "Dynamic panels with MIDAS covariates: Nonlinearity, estimation and fit," *Journal of Econometrics*, 220(2), 589–605.

Kim, H. H., and N. R. Swanson (2014): "Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence," *Journal of Econometrics*, 178, 352–367.

McCracken, M. W., and S. Ng (2016): "FRED-MD: A monthly database for macroeconomic research," *Journal of Business and Economic Statistics*, 34(4), 574–589.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010): "Out-of-sample equity premium prediction: Combination forecasts and links to the real economy," *The Review of Financial Studies*, 23(2), 821–862.

Zou, H., T. Hastie, and R. Tibshirani (2007): "On the "degrees of freedom" of the LASSO," *Annals of Statistics*, 35(5), 2173–2192.