International Journal of Forecasting 32 (2016) 10–19

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Predicting Finnish economic activity using firm-level data

Paolo Fornaro

University of Helsinki, Economicum Arkadiankatu 7, Helsinki 00530, Finland

ARTICLE INFO

Keywords:
Firm-level data
Forecasting
Factor model
Real-time data
Large datasets

ABSTRACT

In this paper, we compute flash estimates of Finnish monthly economic activity using firm-level data. We use a two-step procedure where the common factors extracted from the firm-level data are subsequently used as predictors in nowcasting regressions. The results show that large firm-level datasets are useful for predicting aggregate economic activity in a timely fashion. The proposed factor-based nowcasting model leads to a superior out-of-sample nowcasting performance relative to the benchmark autoregressive model, even for early nowcasts. Moreover, we find that the quarterly GDP flash estimates that we construct provide a useful real-time alternative to the current official estimates, without any substantial loss of nowcasting accuracy.

1. Introduction

Statistical agencies, central banks, and numerous public and private entities collect hundreds, if not thousands, of economic time series every year. This ever-growing amount of data has helped policymakers and researchers with key activities such as forecasting, evaluating the performances of economic models, and designing fiscal and monetary policies. Unfortunately, this wealth of data is not matched by a high degree of timeliness. Most notably, variables measuring economic activity are generally published with long lags. For example, the first estimates of the US and UK quarterly GDP are published four weeks after the end of each quarter, while the lag is usually six weeks for the Euro area (see Banbura, Giannone, & Reichlin, 2011).

In recent years, this problem of the timeliness of data releases has been addressed in the literature on nowcasting models and coincident economic indicators (for the latter, see e.g. Altissimo, Cristadoro, Forni, Lippi, & Veronese, 2010; Stock & Watson, 1989). Nowcasting methods have been applied chiefly in the prediction of low frequency data, and quarterly data in particular, by exploiting the release of monthly data (see, e.g., Aastveit & Trovik, 2014; Banbura et al., 2011; Evans, 2005; Giannone, Reichlin, & Small, 2008). In these papers, the focus has been on the creation of early estimates of the quarterly GDP growth, which are updated as new information is released. These revisions are analyzed by checking the contributions of the news carried by additional data. Most of the nowcasting papers are interested in quarterly variables, though Modugno (2013) and Proietti (2011) focus on computing monthly nowcasts of GDP. Aruoba, Diebold, and Scotti (2009) propose a real-time economic activity indicator that is built on data observed at mixed frequencies, including daily data. Recent examples of nowcasting applications are those of Camacho and Garcia-Serrador (2014) and Camacho and Perez-Quiros (2010), who use a single-index dynamic factor model based on both real and financial indicators, and Foroni and Marcellino (2014), who apply various different approaches (bridge equations, state space and mixed data sampling models) to the nowcasting of Euro area GDP components. Finally, a recent survey on nowcasting with parsimonious mixed-frequency methods is provided by Camacho, Perez-Quiros, and Poncela (2013).

The novel idea introduced in this study is to exploit the information contained in large firm-level datasets in

E-mail address: paolo.fornaro@helsinki.fi.

http://dx.doi.org/10.1016/j.ijforecast.2015.04.002
0169-2070/

order to compute early estimates of economic activity. In particular, we compute nowcasts of the Finnish monthly economic activity indicator, the Trend Indicator of Output (TIO), using a two-step procedure. In the first step, we extract common factors from a large firm-level dataset of turnovers, then in the second step we use these common factors as predictors for nowcasting regressions. The estimates constructed for TIO are also used subsequently to compute early figures of the Finnish quarterly GDP.

This paper has various points in common with each of the aforementioned strands of the literature, but also exhibits substantial differences. In particular, using the factor model of Stock and Watson (2002a,b), we exploit the information contained in large datasets for predicting economic activity. However, we do not formulate a state space model, as is common in the nowcasting literature. Even though the datasets that we use contain the jagged edges (missing values at the end of the sample due to differences in publication times) and missing value problems that are typical of the nowcasting literature, we do not have to deal with mixed frequency data because we focus only on monthly variables and estimate the quarterly GDP from the TIO figures directly.

Another key distinction from the previous literature is that we effectively estimate the economic activity of recent months, reducing the publication lag of TIO figures, without attempting to compute current values of the TIO, based on higher frequency (say weekly) data. Finally, and most importantly, the focus is shifted from the use of public data releases to the use of data available to the statistical agency, namely monthly turnovers data. Indeed, the use of such a disaggregated dataset for nowcasting is the key contribution of this paper to the literature. Of course, this dataset reflects only a (timely) part of the total information set available to Statistics Finland at the time of TIO publication. Factor models are optimal in this scenario because they are able to summarize the important information contained in the data, even though the latter may be incomplete.

In this study, we concentrate on firm-level turnovers only. This is because we want to focus on the information carried by highly disaggregated data for predicting aggregate figures. This is the main contribution of this paper: to the best of our knowledge, no previous paper has used such a disaggregated dataset to nowcast aggregate economic activity in the literature on nowcasting and factor models. Instead, different authors have concentrated on sectoral- or regional-level data (see Banbura et al., 2011; Martinsen, Ravazzolo, & Wulfsberg, 2014). Matheson, Mitchell, and Silverstone (2010) and Mitchell, Smith, and Weale (2013) use firm-level qualitative surveys to predict economic activity and manufacturing; however, we want to stress the fact that we use what the literature refers to as 'hard' data, not qualitative surveys. Alessi, Barigozzi, and Capasso (2013) apply dynamic factor models to firm-level data, but their focus is different from ours. They are interested in studying the dynamics of the business cycles, and have more of a descriptive approach. The dataset that they use is obtained from COMPUSTAT, and the data are quarterly, while we use monthly data. Finally, their analysis focuses on a single data vintage, while in our application we create a series of datasets to replicate the accumulation of information collected by Statistics Finland.

Another, and more subtle, novelty presented in this paper is the use of the regularized expectation maximization (EM) algorithm presented by Josse and Husson (2012a). This method corrects the usual EM estimation of the factors by reducing the risk of overfitting, by taking into account the presence of many missing observations in the factor extraction and in the missing value imputation.

We find that the nowcasts based on the factors extracted from the turnover dataset perform better than the autoregressive benchmark[1] for all periods except for the estimates computed five days after the end of the reference month. Moreover, the mean absolute percentage errors of the nowcasts are not far from the average revisions made by Statistics Finland. This is an encouraging result, in light of the actual implementability of the method. Finally, we find that using the factor nowcasts of TIO in the computation of the quarterly GDP allows us to reduce the publication lag relative to the current official flash GDP estimates published by Statistics Finland, without any loss of nowcasting accuracy.

The remainder of the paper is structured as follows. In Section 2, we present the two-stage statistical model that is employed for constructing nowcasts of TIO. In Section 3, we describe the data, and in particular, the way in which we simulate the accumulation of data over time. The empirical results are presented in Section 4. Finally, Section 5 concludes.

2. Model

In this study, the nowcasting model employed consists of two stages. In the first stage, we extract common factors from a large dataset of firm-level turnovers (Section 2.1). Once the factors have been extracted, they are used in a nowcasting regression (Section 2.2) to construct nowcasts of the variable of interest, namely the monthly year-on-year growth of Finnish economic activity, in this study.

2.1. Factor extraction

The factors are computed as in the factor model of Stock and Watson (2002b). There are multiple reasons for this choice. The datasets that we use to compute the TIO estimates are very large, with the original dataset including over 2000 firms. However, we drop many firms in order to achieve a balanced dataset in the first sample period, ultimately leaving us with 579 firms, which is still a large sample. Hence, we need a model that can handle such a large cross-section but is still computationally feasible to estimate. While the model of Banbura and Modugno (2014) can also handle various data problems and is used widely in the nowcasting literature, it is computationally too demanding for this application. Stock and Watson (2002b)

[1] The nowcasting performance has also been tested against those of a random walk and the TRAMO-SEATS procedure proposed by Gomez and Maravall (2001). The results are similar to those obtained relative to the autoregressive model benchmark, and are available upon request.

have a large dataset that follows a factor model with r latent factors included in F_t. Defining X_t now as the dataset that contains N time series of the growth rates (year-on-year) of firm-level turnovers at time t, we can write their factor model as

$$X_t = \Lambda F_t + e_t, \tag{1}$$

where Λ is the matrix of factor loadings and e_t is the $N \times 1$ vector of idiosyncratic components. The idiosyncratic components are allowed to be both (weakly) serially and cross-sectionally correlated, making this model resemble the approximate factor model of Chamberlain and Rothschild (1983). Given the novelty of our dataset in a nowcasting application, we check for the correlation structure of the idiosyncratic components in Appendix A.2. The factors are estimated by principal components, i.e., $\widehat{F}_t$ is given by the eigenvectors that correspond to the largest eigenvalues of the $T \times T$ matrix XX', where $X = [X_1', \ldots, X_T']'$. This is a handy procedure computationally, because we do not have to deal with very large matrices in the estimation, in spite of the very large cross-section of firms.

A common feature of the datasets used in nowcasting exercises, like that in this paper, is the presence of jagged edges and missing values. The basic principal component estimation requires a balanced dataset (i.e., all of the time series should be of the same length and without missing values). In this study, we deal with the missing values problem in two different ways. The first method simply involves creating a balanced dataset by taking a subset of the variables from the original dataset. This means that we do not have to perform missing value imputation, with the associated estimation errors and computational intensity; however, we do have to give up a part of the original dataset, at least for the very early estimates. We refer to this methodology later on as a balanced method.

As an alternative procedure, we use the regularized iterative principal component analysis (PCA) algorithm (see Josse & Husson, 2012a, for details). This method is preferred to the simple EM iterative PCA presented by Stock and Watson (2002a) because it is aimed at datasets with many missing values, as is the case with the data to be analyzed in Sections 3 and 4. Moreover, the regularized iterative PCA method performs better with respect to the overfitting problem, due to the fact that the regularized iterative principal component algorithm shrinks the weight of the principal components in the missing value imputation, for datasets with large numbers of missing values.

The simple EM-PCA algorithm consists of three steps. In the first step, we impute some initial guess for the missing values. One possibility is to impute the mean of each variable, while Stock and Watson (2002a) suggest the use of an initial balanced dataset for computing the first estimate of the factors. In the second step, we use the estimated factors to impute the missing data, following

$$\widehat{X}_{tk} = \hat{\mu}_k + \sum_{s=1}^{S} \widehat{F}_{ts} \widehat{\Lambda}_{ks}, \tag{2}$$

where $\widehat{X}_{tk}$ is a missing value at time t for the variable k, $\hat{\mu}_k$ is its mean, and S is the chosen number of factors. In the last step, we estimate the factors from the dataset with the imputed values. We iterate these three steps until we reach convergence (for a formal proof, see Dempster, Laird, & Rubin, 1977).

The basic idea behind this regularized PCA algorithm is that if there is a lot of noise in the data, or, equivalently, if the structure of the dataset is too weak (for example, lots of missing values), the algorithm weights the principal component imputation ($\sum_{s=1}^{S} \widehat{F}_{ts} \widehat{\Lambda}_{ks}$ in Eq. (2)) less, and tends to impute the simple mean of the variable (μ_k). If there is little noise in the data, then this algorithm reduces to the simple EM algorithm of Stock and Watson (2002a). More formally, the regularized PCA algorithm shrinks the principal component part of the imputation step, to get

$$\widehat{X}_{tk} = \hat{\mu}_k + \sum_{s=1}^{S} \left(\frac{\hat{\lambda}_s - \hat{\sigma}^2}{\hat{\lambda}_s} \right) \widehat{F}_{ts} \widehat{\Lambda}_{ks}, \tag{3}$$

where $\hat{\lambda}_s$ is the s singular value of the matrix X and $\sigma^2 = \frac{1}{K-S} \sum_{s=S+1}^{K} \hat{\lambda}_s$, which can be interpreted as the amount of noise in the data.

The trade-off between the balanced method and the iterative PCA method stands out from the fact that we do not have to go through the missing values imputation process in the balanced method. This is a time consuming process, and, more importantly, may cause bad predictions of the missing values, which could create problems for the factor extraction, and therefore unnecessary bias in the second stage (nowcasting) of our model. On the other hand, the iterative PCA has the advantage that it provides an efficient way to use all of the firms included in the dataset.

2.2. Nowcasting model

In the second stage of our model, we use the estimated factors as predictors in the nowcasting model

$$y_t = \beta_\nu \widehat{F}_{t|\nu} + \epsilon_{t|\nu}, \tag{4}$$

where y_t measures the monthly economic activity, with t being the reference month we are interested in, $\epsilon_{t|\nu}$ is the nowcasting error, and ν is the period in which we compute our nowcast (i.e., the number of days after the end of the reference period that we compute the estimate). In our application, we estimate Eq. (4) nine times for each period, namely at $\nu = \{5, 10, 15, 20\}$ up to $\nu = 45$ days after the end of the reference month (see Section 3 for details). We do not compute factor estimates after $\nu = 45$ because the economic activity indicators have usually been released by that time. Nowcasts that minimize the mean squared error are constructed as $\hat{y}_{t|\nu} = \hat{\beta}_\nu \widehat{F}_{t|\nu}$, where $\hat{y}_{t|\nu}$ denotes the predicted value at time ν and the parameters β_ν are estimated by ordinary least squares (OLS).

One important issue in the estimation process stems from factor selection, namely how many factors should be included in $F_{t|\nu}$. For the balanced method, factor selection can be based on information criteria, such as the Bayesian Information Criterion (BIC) or the factor-based regression criterion suggested by Groen and Kapetanios (2013). As a robustness check, we also compute nowcasts based on 10 factors and check the out-of-sample performances of the various models. The estimation of the number of factors

is an even more delicate matter when we deal with missing value replacement (see Section 2.1). Josse and Husson (2012b) provide an algorithm that estimates the optimal number of principal components for a given dataset with missing values.

3. Data description

The variable that we are interested in for this study is the year-on-year growth rate of the Trend Indicator of Output (TIO), which measures Finnish economic activity on a monthly basis. The sample period starts in January 1999 and ends in December 2012. In the out-of-sample nowcasting experiment in Section 4, we nowcast TIO starting from January 2006, giving us a total of 84 observations. To allow for the statistical agency applying possible modifications to the indicator, we do not seasonally adjust the original TIO series. However, taking year-on-year growth rates should remove possible seasonal components. We also follow this strategy for the firm turnovers. To check for the absence of seasonality in our data, we regress the firm turnovers and the TIO onto a constant and a set of seasonal dummies. We find that the p-value associated with the F-statistic of these regressions is less than 5% only twice, indicating that seasonal effects are not important in our turnover dataset. A similar result is found for the TIO and the estimated factors.

The TIO is currently released by Statistics Finland with two different time schedules, depending on the reference month. For the first two months of a given quarter, the TIO is released 65 days after the end of the reference month. For the last month of a given quarter, it is published 45 days after the end of the reference period.[2] The TIO is revised after its first publication, and these revisions reflect both changes in the source of data and revisions due to benchmarking. The data sources are split between the price development and value data, which are aggregated to a 2-digit level. The primary sources of data for private manufacturing are the preliminary turnover indices (which are accumulated quickly), while the main sources for the public sector are preliminary wages and salaries.

In the Finnish system of national accounts, a flash estimate of the quarterly GDP growth is published 45 days after the end of the reference quarter, based on the TIO. Fig. 1 depicts the time series of the TIO (year-on-year percentage changes) over the period examined in this study. The TIO is deflated using average prices of the year before, using the year 2000 as a reference year. The turnover data, on the other hand, are not price adjusted. This feature is handled well by factor models, which are able to separate the main co-movements in the data and can be compared to the presence of nominal variables in a large dataset such as that used by Stock and Watson (2002a).

The Finnish economy expanded during the period from January 1999 to December 2007, with a mean year-on-year growth of 3.6%. Since the start of the recent recession in 2008, however, Finland has faced a dramatic drop in output (a mean growth rate of −0.7%).

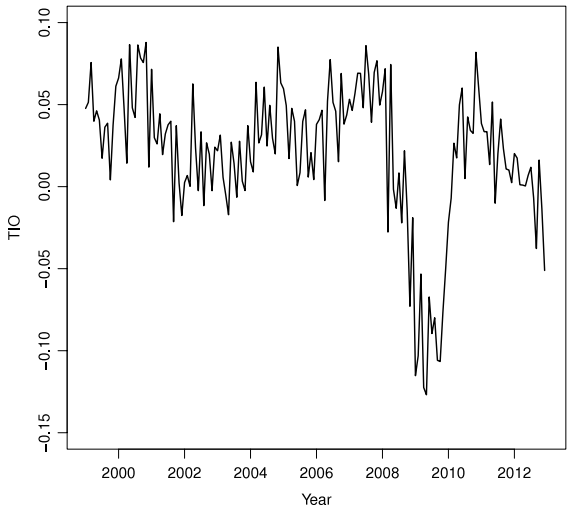


Fig. 1. Plots of the TIO year-on-year percentage changes during the sample period.

A major contribution of this study is to use firm-level data in factor estimation for nowcasting purposes. Due to its timeliness, firm-level turnover data appears to be an interesting alternative to the datasets previously used in factor extraction (see e.g. Giannone et al., 2008). These data are accumulated right after the end of the reference month, and the date on which a firm sends its data to Statistics Finland is documented carefully and collected in a dataset. Thanks to these reports, we can replicate the real-time data environment closely. However, due to confidentiality issues, this firm-level dataset is not available publicly.

In our nowcasting experiment, we simulate the data accumulation process by creating different real-time datasets of the year-on-year growth rates of turnover indices available at different periods. For each month, we create nine different datasets (i.e., nine different values of ν in Eq. (4)), corresponding to turnover indices available at $t|5$, $t|10$, and so on. For example, when we estimate TIO in December 2009 at $\nu = 20$, we base our estimation on turnovers available by January 20th, 2010, and we use only turnovers of private firms as predictors in the dataset.

While it is true that some additional data could be useful, we want to extract and isolate as much as possible the ability of this particular firm-level dataset to give very early signals of the TIO. Given the originality of this dataset for our nowcasting application, it is useful to examine its predictive power in the most straightforward way possible; adding more predictive variables would simply complicate the analysis. Moreover, focusing on turnover indices allows a very precise replication of the data accumulation, which becomes much more cumbersome when additional data sources are examined.

The original turnovers dataset contains more than 2000 firms, but many of these time series present extremely high numbers of missing values. Because we want to compute nowcasts starting from the beginning of 2006 and to start the estimation period as early as possible, we exclude several firms from the dataset. We keep the firms that had started reporting by 1999 and reported until at least the

[2] A calendar of future releases can be found at http://tilastokeskus.fi/til/ktkk/tjulk_en.html.

end of 2005. This gives an initial balanced dataset for the estimation. The remaining dataset includes 579 firms for the initial sample period. Once we drop the firms that have missing values in the initial sample period, the volume of turnovers of the remaining firms amounts to 45% of the total turnovers in the original dataset, increasing to 64% of total turnovers by the end of 2005, and to 77% by the end of 2012. While the information loss seems quite large at the beginning of the sample, the later periods seem to contain a large fraction of the total turnovers in the original dataset.

In Appendix A.1, we report additional information about the data accumulation process, including, for example, the average percentage of firms reporting by ν days after the end of the reference period, and the plot of the cumulative eigenvalues for the turnovers dataset. These statistics are useful for analyzing how much the data accumulation affects our estimates and how much the information contained in this large disaggregated dataset can be squeezed into a few factors.

4. Empirical results

We compute nowcasts by following the methods described in Section 2. The initial in-sample period goes from January 1999 to December 2005, whereas the nowcasting period starts from January 2006. We re-estimate the model forward using an expanding window up to December 2012. We start our analysis of the empirical results by having a look at the plots of the nowcasts against the original series. While this is an informal method of analyzing the results, even a visual inspection of the nowcasts can give important insights into their performances.

In Fig. 2, we show the nowcast for the TIO for the model using the factors selected by the BIC. We report only the nowcasts obtained using the BIC in this section, as the criterion of Groen and Kapetanios (2013) led to similar results. We compare prediction performances based on the root mean square forecast error (RMSFE), using an AR(p) model as the benchmark, where the lag length p is selected based on the BIC. Moreover, we also compute the mean absolute percentage error of the predictions, to shed some light on the actual applicability of the method.

Fig. 2 depicts the nowcast performance using the balanced method with the factors selected based on the BIC. We immediately see that at $\nu = 5$, i.e., five days after the end of the reference period, the nowcasts are pretty inaccurate. Even though the nowcasts follow the overall trend of the series, there are some large deviations, such as at the end of 2009 and 2011. Remember that in the case of $\nu = 5$, the nowcasts are based on turnovers from a very small set of firms. Moving even to $\nu = 10$ or $\nu = 15$, we have a fairly large improvement: there seem to be many fewer implausible spikes, and the nowcasts seem to track the original series much better.

Another interesting feature is that there are no visible improvements from going more than 20 days after the end of the reference period. This indicates that the selection of $\nu = 20$ might be optimal for the factor model in Eq. (4) in terms of the tradeoff between timeliness and the accuracy of the nowcasts. This selection is able to pick up the most interesting co-movements in the turnover dataset, and it also appears that any further increase in the accuracy of the nowcast might require the model to be augmented with some additional predictive variables. This impression gained from the plots is confirmed in Table 1, where the root mean squared forecast errors (defined in Eq. (5)) do not change substantially after $\nu = 25$.

Next, in Fig. 3, we depict plots of the nowcasts based on the factors extracted using the regularized EM algorithm. As above, the factors used in the nowcasting regression are selected using the BIC. This allows us to perform missing value imputation, using the relative prediction error, but we can use larger datasets even at earlier times.

We see immediately that there is a substantial degree of smoothing in the nowcasts computed at $t|5$, even though, similarly to Fig. 2, they remain inaccurate. Looking at the root mean squared error results, we see that the balanced method provide a better alternative for these very early estimates, even though neither method is able to beat the benchmark AR model for $\nu = 5$. Overall, nowcasts based on the regularized iterative PCA seem to perform well, being able to predict at the very end of the sample as well, which is something that the balanced method seems to have difficulties with.

Even though graphs can give a general indication of how well the methods perform, we still need to use some numerical evaluation criteria for judging the out-of-sample performances of the models at hand. We use two different measures: the root mean square forecast error and the absolute percentage error. The former is defined as

$$\text{RMSFE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_{t|\nu})^2}. \tag{5}$$

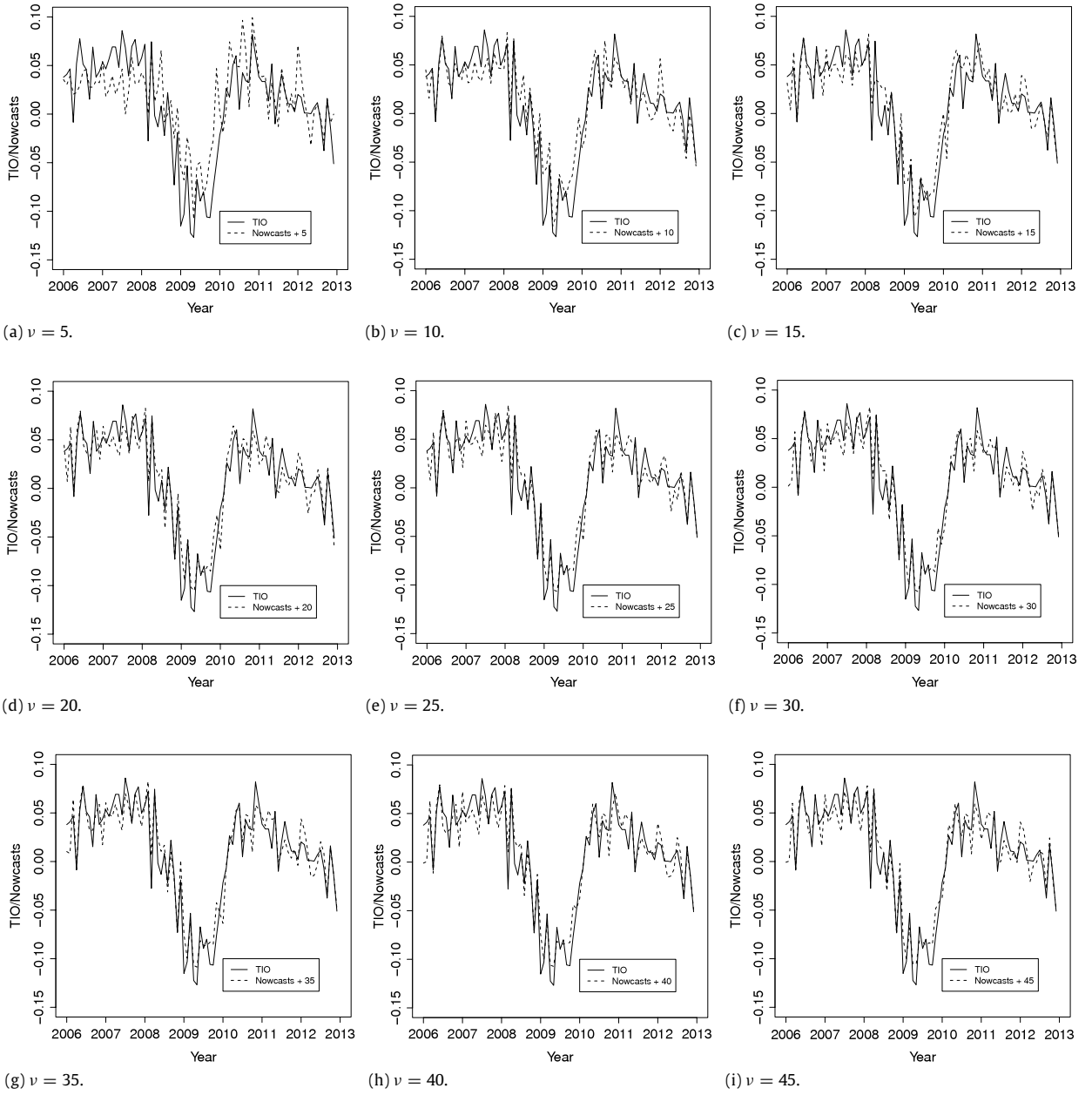
In our tables, we report relative RMSFEs, which are computed relative to the RMSFE of the benchmark model. Thus, a value below one indicates that our nowcasting model provides better nowcasts than the benchmark models. The other measure, the mean absolute percentage error, given by

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{(y_t - \hat{y}_{t|\nu})}{y_t}\right|,$$

indicates how far our estimates are from the true value on average, thus giving a good indication of the method's performance in the light of a practical implementation. Moreover, we rely on the Diebold and Mariano (1995) test when comparing the predictive accuracies of two non-nested nowcasting models. Throughout this analysis, we use the AR(p) model as our benchmark model.

In Table 1, we report the relative RMSFE for the balanced (Bal.) and EM (EM) methods using 10 factors and factors selected by the BIC.

According to Table 1, it seems that the methods proposed here are able to beat the benchmark AR(p) model for most of the periods ν. The relative RMSFEs are consistently below unity. Only the nowcasts performed five days ($\nu = 5$) after the end of the reference period are worse than the benchmark. The nowcasts based on the EM algorithm perform better at $t|5$, but seem to offer only a moderate advantage over the balanced method. Another



(a) $\nu = 5$. (b) $\nu = 10$. (c) $\nu = 15$.

(d) $\nu = 20$. (e) $\nu = 25$. (f) $\nu = 30$.

(g) $\nu = 35$. (h) $\nu = 40$. (i) $\nu = 45$.

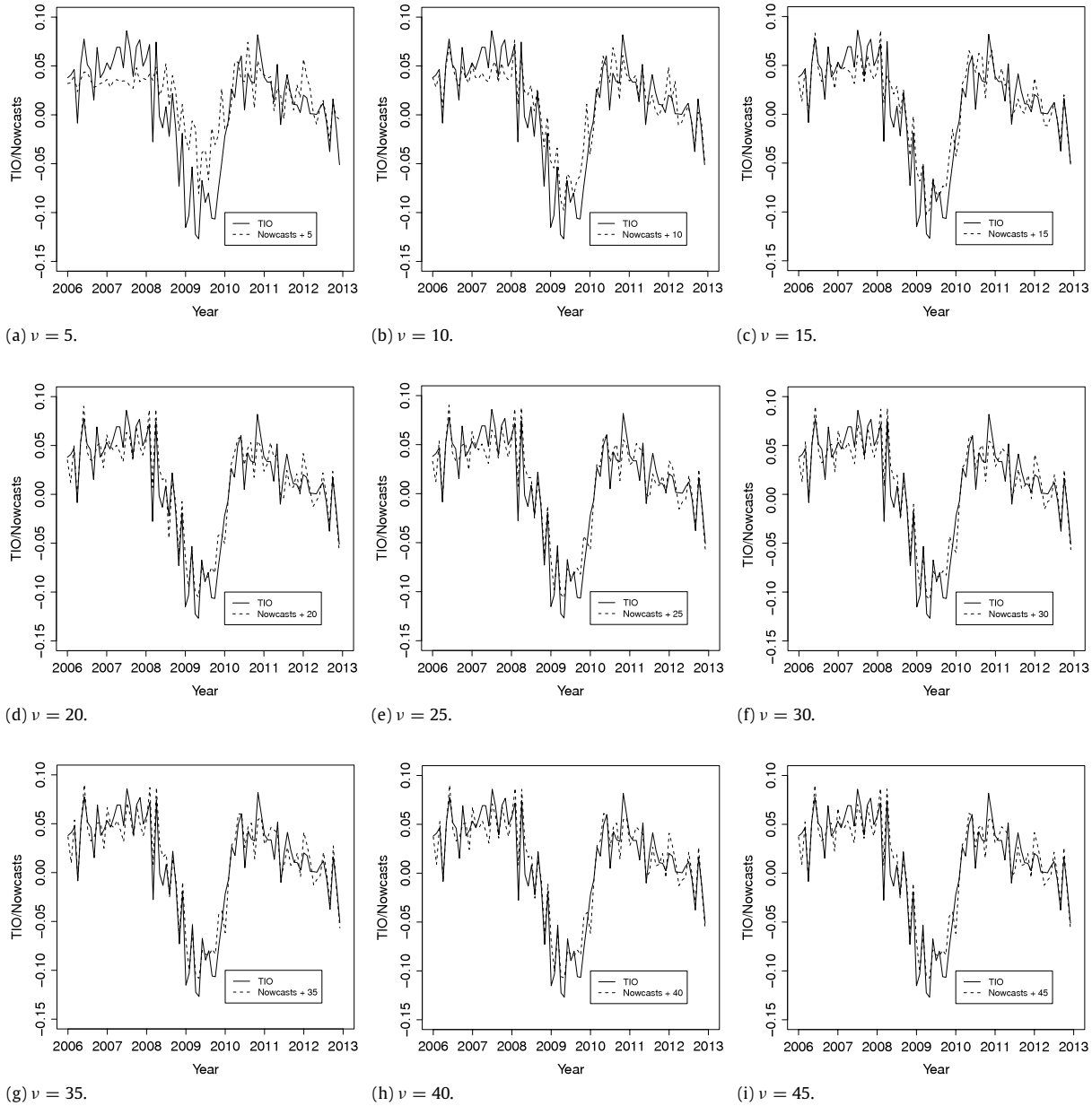
Fig. 2. Nowcasts computed with the balanced method at $t|\nu$.

interesting aspect is the fact that the predictive performance does not improve much after $\nu = 20$. For our nowcasting application, the principal components are able to estimate the important underlying factors by using only a subset of the firms, without needing the complete dataset. This overall superiority in predictive performance is also confirmed by the results of the Diebold and Mariano (1995) test.

It is also important to have an idea of how much our predictions deviate from the actual (revised) values of TIO, in order to evaluate how well the models perform in practice. Table 2 reports the mean absolute percentage errors for the TIO year-on-year percentage changes.

It turns out that the balanced-method-based predictions perform better at $\nu = 5$, while the EM-based nowcasts are slightly better later on. Also, the more parsimonious models seem to create worse estimates than the models with more factors included in the nowcasting model in Eq. (4).

Overall, in Table 2, looking at the EM-based nowcasts, the usual percentage deviation from the actual revised TIO value is 1.3%. The usual revision done by Statistics Finland is around 0.9%, so our estimates are somewhat worse than those made by Statistic Finland (we get around a 44% loss in accuracy). This is to be expected, as the Statistics Finland revisions are based on the actual figures, which



(a) $\nu = 5$. (b) $\nu = 10$. (c) $\nu = 15$.

(d) $\nu = 20$. (e) $\nu = 25$. (f) $\nu = 30$.

(g) $\nu = 35$. (h) $\nu = 40$. (i) $\nu = 45$.

Fig. 3. Nowcasts computed using the EM method at $t|\nu$.

have substantial publication lags and are based on a much wider dataset. However, the nowcast errors do not differ dramatically from the revisions of the initial estimates computed by Statistics Finland, meaning that, taking the reduction in the publication lag into account, our method provides an attractive alternative to the method currently used. Ultimately, the question is how much the statistical agency values smaller revisions relative to having a more timely indicator. As we will see, the larger deviations from the revised values of the TIO are not reflected in the figures for the quarterly GDP.

So far, we have focused on the nowcasts of the monthly TIO; however, a very interesting application of this methodology lies in the prediction of quarterly GDP. In particular, we can use the nowcasts constructed for the TIO to compute early estimates of the GDP, with shorter publication lags than the current GDP flash estimate published by Statistics Finland around 45 days after the end of the reference quarter. Using the method presented in this paper, we can shorten the publication lag considerably. One possibility is to estimate the quarterly GDP using the classical TIO measurement for the first two months of a given quarter and use the factor-based nowcast for the last month. Even though we have seen that the nowcasting performance does not improve greatly after $\nu = 20$, the predictions get slightly better at $\nu = 25$. Moreover, Statistics Finland publication procedures do not realistically allow a release of

Table 1
Relative RMSFEs for nowcasts of TIO in percentage changes, where the AR model is used as a benchmark.

ν	BIC factor (Bal.)	10 factor (Bal.)	BIC factor (EM)	10 factor (EM)
5	1.05	1.06	1.15	1.17
10	0.65***	0.65***	0.73***	0.77***
15	0.68***	0.65***	0.65***	0.66***
20	0.59***	0.60***	0.55***	0.57***
25	0.55***	0.59***	0.54***	0.57***
30	0.57***	0.58***	0.54***	0.56***
35	0.57***	0.58***	0.54***	0.56***
40	0.60***	0.58***	0.54***	0.56***
45	0.58***	0.57***	0.54***	0.56***

*** Indicates rejection of the hypothesis of equal predictive ability at the 1% statistical significance level.

Table 2
Mean absolute percentage errors for nowcasts of TIO in year-on-year percentage changes.

ν	BIC factor (Bal.)	10 factor (Bal.)	BIC factor (EM)	10 factor (EM)
5	0.024	0.025	0.026	0.026
10	0.015	0.015	0.017	0.017
15	0.016	0.015	0.015	0.015
20	0.014	0.014	0.013	0.014
25	0.013	0.015	0.013	0.014
30	0.013	0.014	0.013	0.014
35	0.013	0.014	0.013	0.014
40	0.014	0.014	0.013	0.014
45	0.014	0.014	0.013	0.014

Table 3
Mean absolute percentage errors for the obtained nowcast and flash estimates of quarterly GDP (year-on-year percentage changes) for different sample periods.

ν	2006–2012	2008–2012	2010–2012	2012
$t-25$ factor estimates	0.0059	0.011	0.009	0.004
$t-45$ flash estimates	0.0054	0.008	0.007	0.006

GDP before 25 days after the end of the reference quarter. For these reasons, we use the nowcasts obtained at $\nu = 25$.

In Table 3, we report the mean absolute percentage errors (relative to the revised GDP figure) obtained by predicting quarterly GDP year-on-year changes using this method. The factor model employed is based on the EM method, and we report results based on $t|25$ estimates. The number of factors is selected using the BIC.

The results presented in Table 3 are very encouraging. It seems that we can shorten the publication lag considerably without any major increase in the revision error. In particular, in the six years between 2006 and 2012, the nowcasting errors of the factor model and the current flash estimates are essentially equal, while the factor method manages to beat the current estimates for the year 2012. Based on these results, we conclude that nowcasts based on the proposed factor model provide a competitive method for nowcasting GDP growth.

5. Conclusions

In this study, we have used a large dataset of Finnish firm-level turnovers to compute factors which are in turn included in a predictive regression for nowcasting monthly economic activity. We compute the factors using two methods. In the first method, we simply eliminate the firms that present jagged edges or missing values, thus ensuring that the turnover dataset is balanced, and use a simple principal component estimator to extract the factors. We call this routine a balanced method. In our other method, we perform missing value imputation using the factor model and the regularized EM algorithm proposed by Josse and Husson (2012a). This method allows us to use all of the firms in the dataset, but is also computationally more intensive than the balanced method.

We find that both of these methods beat the benchmark nowcasts based on the AR model for all estimation periods except for very early periods close to the end of the month that we want to nowcast. We also find that the EM method does provide better nowcasts than the balanced method, but the improvement is not very large. Finally, we find that the factor-based nowcasts provide a competitive alternative to the current flash estimates of quarterly GDP year-on-year growth. In particular, we see that the nowcasts computed using this method allow substantially shorter publication lags (in our case, a 20-day reduction in the publication lag). Overall, our main finding is that the factors extracted from a large micro dataset are useful for predicting economic activity.

There are several possible extensions to this paper. The most obvious one is to expand the initial cross-section of variables used in the factor extraction. Along with firm-level turnovers, we could also include macroeconomic and financial variables in our nowcasting model. Moreover, the factors and the TIO estimates obtained in this exercise could be used in a wider nowcasting application. Very early nowcasts can be produced based on surveys and financial variables, but as time goes on, we can also add the TIO estimates as indicators in the nowcasting equations. In addition, the nowcasting regression used could be extended, for example, by adding lags of the dependent variable or the constructed factors. We could also use models which take factor dynamics into account in the factor estimation (see e.g. Doz, Giannone, & Reichlin, 2011).

Acknowledgments

I would like to thank Esther Ruiz, the associate editor, and two anonymous referees for their thoughtful comments and suggestions. I am especially grateful to Faiz Alsuhail, Samu Hakala, Pasi Koikkalainen, Ville Koskinen, Henri Luomaranta and the rest of the staff at Statistics Finland for the invaluable help with the data and comments. Moreover, I wish to thank my supervisors Henri Nyberg and Antti Ripatti, and the other participants at University of Helsinki seminars, at the DIW Macroeconometric Workshop 2013 and at the 25th $(EC)^2$ Conference in Barcelona for their constructive comments. This paper was written during my internship at Statistics Finland.

Table A.4
Accumulation of turnover data by ν days after the end of the reference month.

ν	5	10	15	20	25	30	35	40	45
Number of firms reporting	35	125	262	389	432	454	460	465	468
% of firms reporting	7	26	56	83	91	96	97	98	100
% of turnovers reported	7	25	57	84	92	97	97	98	100

Table A.5
Average autocorrelations and cross-sectional correlations of the idiosyncratic components for different firms.

Firms	Whole sample	Top 10%	Bottom 10%
Autocorrelation	0.11	0.10	0.11
Cross-correlation	0.002	0.02	0.01

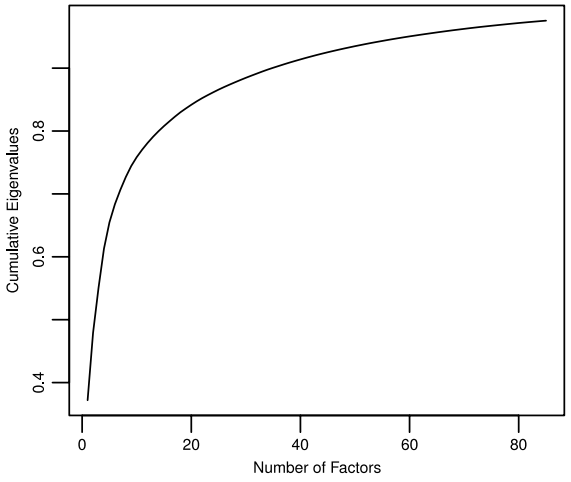


Fig. A.4. Cumulative eigenvalues plot.

Appendix

A.1. Data accumulation

One of the nice features of the dataset used in this study relates to the possibility of tracking the data accumulation obtained by the official statistical agencies such as Statistics Finland. It is interesting to see how the data accumulation evolves over time, reflecting the dynamics of the information available to the data producer. In Table A.4, we report the average numbers and percentages of firms sending their turnover data to Statistics Finland at different points in time after the end of a given period (denoted by ν). Note that the percentage is calculated with respect to the number of firms that have reported by $\nu = 45$. This decision causes the average number of firms reporting by $\nu = 45$ to be less than the total number of firms present in the dataset (579), even though we take the percentage of firms reporting by $\nu = 45$ to be 100%. Many firms send their data more than 45 days after the end of the reference month. We also include the percentage of total turnovers reported by a given date, in order to check whether there is any relationship between firm sizes and the timeliness of their reports.

The accumulation of the data seems to become very slow once it is more than $\nu = 20$ or $\nu = 25$ days after the end of the reference month. This is also reflected in the fact that the nowcasting performance does not improve much after $\nu = 20$ (see Section 4). Moreover, the percentage of firms reporting and the percentage of turnovers accumulated are very similar. This indicates that there is no specific pattern as to which kinds of firms report their turnovers first. If the largest firms sent their turnovers first, then we would find that the turnover accumulation was faster than the percentage of firms reporting.

A.2. Factor properties

Given the novelty of our dataset, we are interested in seeing whether some of the basic assumptions of the factor model described in Section 2.1 are met. In particular, we want to check that the idiosyncratic components do not present strong serial and cross-sectional correlations. In Table A.5, we report average absolute first-order autocorrelations and cross-correlations for the whole sample of firms. Moreover, we divide the sample of firms by the size (calculated as the time average of the ratio of a firm turnover to the total turnovers) and compute the correlations for the bottom and top 10th percentiles.

From Table A.5, we see that the idiosyncratic errors do not display large serial correlations. Furthermore, according to the Monte Carlo experiment presented by Stock and Watson (2002b), factor estimation with principal components is effective even with moderately autocorrelated errors, noting that the example that the authors use involves correlations higher than those reported in Table A.5. Thus, it seems that the principal component estimation using our dataset should not generate problems in terms of error dependencies.

Another interesting question in relation to this highly disaggregated dataset is how much information can be squeezed into constructed factors. To shed some light on this matter, Fig. A.4 reports the plot of cumulative eigenvalues for the dataset of turnovers in December 2012 of firms reporting by January 31st (that is, the last and most extensive vintage available).

This plot, together with Table A.6, gives us a rough idea of how much of the variance in the turnover dataset is explained by the common factors. Even though the cumulative eigenvalues do not increase greatly after the fourth factor, we find that a rich model, with more than 20 factors in the nowcasting model in Eq. (4), performs well.

Table A.6
Variance of the firm-level dataset explained by common factors.

Number of factors	1	2	3	4	5	6	7	8	9	10
	0.37	0.47	0.55	0.61	0.65	0.68	0.70	0.72	0.74	0.75
Number of factors	11	12	13	14	15	16	17	18	19	20
	0.77	0.78	0.79	0.79	0.80	0.81	0.82	0.82	0.83	0.84

References

Aastveit, K. A., & Trovik, T. G. (2014). Estimating the output gap in real time: A factor model approach. *Quarterly Review of Economics and Finance*, *54*(2), 180–193.

Alessi, L., Barigozzi, M., & Capasso, M. (2013). The common component of firm growth. *Structural Change and Economic Dynamics*, *26*, 73–82.

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., & Veronese, G. (2010). New Eurocoin: Tracking economic growth in real time. *The Review of Economics and Statistics*, *92*(4), 1024–1034.

Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business and Economic Statistics*, *27*(4), 417–427.

Banbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting. In *Oxford handbook of economic forecasting*. Oxford University Press.

Banbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, *29*(1), 133–160.

Camacho, M., & Garcia-Serrador, A. (2014). The euro-sting revisited: The usefulness of financial indicators to obtain euro area GDP forecasts. *Journal of Forecasting*, *33*(3), 186–197.

Camacho, M., & Perez-Quiros, G. (2010). Introducing the euro-sting: Short-term indicator of euro area growth. *Journal of Applied Econometrics*, *25*(4), 663–694.

Camacho, M., Perez-Quiros, G., & Poncela, P. (2013). Short-term forecasting for empirical economists: A survey of the recently proposed algorithms. *Foundations and Trends in Econometrics*, *6*(2), 101–161.

Chamberlain, G., & Rothschild, M. (1983). Arbitrage, factor structure, and mean–variance analysis on large asset markets. *Econometrica*, *51*(5), 1281–1304.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

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

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Evans, M. D. D. (2005). Where are we now? Real-time estimates of the macroeconomy. *International Journal of Central Banking*, *1*(2), 127–175.

Foroni, C., & Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. *International Journal of Forecasting*, *30*(3), 554–568.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: the real time informational content of macroeconomic data releases. *Journal of Monetary Economics*, *55*(4), 665–676.

Gomez, V., & Maravall, A. (2001). Automatic modelling methods for univariate series. In *A course in time series analysis*. Wiley and Sons.

Groen, J. J. J., & Kapetanios, G. (2013). Model selection criteria for factor-augmented regressions. *Oxford Bulletin of Economics and Statistics*, *75*(1), 37–63.

Josse, J., & Husson, F. (2012a). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, *153*(2), 79–99.

Josse, J., & Husson, F. (2012b). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics and Data Analysis*, *56*(6), 1869–1879.

Martinsen, K., Ravazzolo, F., & Wulfsberg, F. (2014). Forecasting macroeconomic variables using disaggregate survey data. *International Journal of Forecasting*, *30*(1), 65–77.

Matheson, T., Mitchell, J., & Silverstone, B. (2010). Nowcasting and predicting data revisions using panel survey data. *Journal of Forecasting*, *29*(3), 313–330.

Mitchell, J., Smith, R., & Weale, M. (2013). Efficient aggregation of panel qualitative survey data. *Journal of Applied Econometrics*, *28*(4), 580–603.

Modugno, M. (2013). Now-casting inflation using high frequency data. *International Journal of Forecasting*, *29*(4), 664–675.

Proietti, T. (2011). Estimation of common factors under cross-sectional and temporal aggregation constraints. *International Statistical Review*, *79*(3), 455–476.

Stock, J. H., & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. In *NBER macroeconomics annual 1989. Vol. 4* (pp. 351–409). National Bureau of Economic Research, Inc.

Stock, J. H., & Watson, M. W. (2002a). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, *20*(2), 147–162.

Stock, J. H., & Watson, M. W. (2002b). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.

Paolo Fornaro. Born in Torino, Italy, on 24th June 1987, I did my bachelor studies in Bocconi University in Milan, where I completed a Bachelor in International Management and Economics. I subsequently moved to Finland to do my masters studies in economics. I completed this at the University of Helsinki in 2011, with my thesis focusing on forecasting using factor models. Afterward, I continued my studies at the University of Helsinki, where I started a Ph.D. in economics. My main research topics are factor models and forecasting and nowcasting, and I currently collaborate with Statistics Finland on these topics.