Working Papers 2/2013

Predicting Finnish economic activity using firm-level data

Paolo Fornaro

University of Helsinki
Faculy of Social Sciences

Tilastokeskus
Statistikcentralen
Statistics Finland

Inquiries:

Faiz Alsuhail
09 17 341

Principal editor: Faiz Alsuhail

Layout: Hilkka Lehtonen

The publications in the Working Papers series do not represent Statistics Finland’s viewpoints.
The writers are responsible for the contents of the publications.

Quoting is encouraged provided Statistics Finland is acknowledged as the source.

ISSN 2323–1998
= Working Papers
ISBN 978–952–244–460–8

Predicting Finnish economic activity using firm-level data[1]

Paolo Fornaro[2]

University of Helsinki
Faculy of Social Sciences

Abstract

In this paper we compute flash estimates of the Finnish monthly economic activity indicator, using firm-level data. We use a two-step procedure where the extracted common factors from the firm-level data are subsequently as predictors in nowcasting regressions. These factors-based nowcasting models lead to a superior out-of-sample performance compared with the benchmark models including autoregressive and random walk benchmark, even for very early estimates. Moreover we find that quarterly GDP flash estimates based on factor models provide a timelier alternative to the current estimates, without loss in accuracy. We show that large firm-level datasets are useful in predicting aggregate economic activity in a timely fashion.

Keywords: Firm-level Data, Forecasting, Factor Model, Real-Time Data, Large datasets

[1] I would like to thank 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. The results on quarterly GDP are based on calculations by Samu Hakala. Moreover, I wish to thank Henri Nyberg, Antti Ripatti and the other participants at University of Helsinki seminars for their constructive comments. This paper was written during my internship in Statistics Finland.

[2] Department of Political and Economic Studies, and HECER, University of Helsinki. E-mail: paolo.fornaro@helsinki.fi

Contents

1 Introduction

Statistical agencies, central banks, and numerous public and private entities collect hundreds if not thousands of economic series every year. This ever-growing wealth of data has helped policymakers and researchers in key activities such as forecasting, evaluating the performance of economic models and designing fiscal and monetary policies. Unfortunately this wealth of data is not matched with a high degree of timeliness. Most notably, variables measuring economic activity are published with long lags. For example, the first estimates for the US and UK quarterly GDP are published with four weeks after each quarter, while for the Euro Area the lag is usually six weeks (see Banbura et al. 2010).

The problem of the timeliness of data release has been addressed in the recent years in the literature of nowcasting models and coincident economic indicators (see Stock and Watson 1989, Altissimo et al. 2007 and Aruoba et al. 2009).

Nowcasting has been mostly applied in the prediction of low frequency data, in particular quarterly data, by exploiting the releases of monthly data (see e.g. Banbura et al. 2010, Aastveit and Trovik 2008, Evans 2005 and Giannone et al. 2008). Their focus has been to create early estimates of quarterly GDP, which are updated with the release of new information. These revisions are analyzed by checking the contribution of news carried by additional data. Most of the nowcasting papers are interested in quarterly variables, whereas Modugno (2011) and Proietti (2008) are interested in computing monthly nowcasts of GDP.

In this study, the novel idea is to exploit the information contained in a large firm-level dataset 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, whereas in the second step we use these factors in nowcasting regressions. The estimates of TIO are also used to compute early figures of quarterly GDP.

This paper presents points in common with both aforementioned literatures, but presents substantial differences. As in the nowcasting methodology, we exploit large dataset's information to predict economic activity indicators. In particular, we use the factor model by Stock and Watson (2002a; 2002b) but we do not formulate a state-space model as it is common in the nowcasting literature. Even though the datasets we use present jagged edges (missing values in the end of the sample due to different publication times) and missing values problems as in the nowcasting literature, we do not have to deal with mixed frequency data, because we focus on monthly data and we estimate the quarterly GDP directly from the TIO figures. Another key distinction is that we effectively estimate the economic activity of recent months, reducing the lag in the publication of TIO figures, without attempting to compute current values of TIO, based on higher frequency (say weekly) data. Finally, and most importantly, the interest is shifted from the use of public data release to the use of data available to the statistical agency, namely monthly turnovers data. Indeed, the use of such disaggregated dataset in nowcasting is the key contribution of this paper to the literature. 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 data may be incomplete.

We concentrate on firm-level turnovers only. The reason is that we want to focus on the information carried by highly disaggregated data to predict aggregate figures, which is the main contribution of this paper to the research on the field. To the best of our knowledge, there are no previous papers using such a disaggregated dataset to predict aggregate economic activity in the literature on nowcasting and factor models. Instead different authors have concentrated on sectorial or regional level data (see Banbura et al. 2010) and Martinsen et al. 2010). Matheson et al. (2007) and Mitchel and Weale (2005) use firm-level qualitative surveys to predict economic activity and manufacturing. We want to stress the fact that we use what the literature calls 'hard' data and not qualitative surveys. Alessi et al. (2006), apply dynamic factor models to firm-level data, but their focus is very different from ours. They are interested in studying the dynamics of the business cycle and they have more of a descriptive approach. The dataset they use is obtained from COMPUSTAT and the data are quarterly instead of monthly, as in our case. Even more importantly they do not deal with the real-time data accumulation problem. In other words, their data is based on a single vintage, whereas our dataset allows us to track the actual data accumulation faced by Statistics Finland and it is optimal for simulation of a real-time environment.

Another, more subtle, novelty presented in this paper, is the use of the regularized expectation maximization (EM) algorithm presented in Josse and Husson (2012b). 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 values imputation.

We find that nowcasts based on the factors extracted from the turnover datasets perform better than the autoregressive and random walk benchmarks for all the periods except for the estimates computed five days after the end of the reference period. Moreover, the mean absolute percentage errors of the nowcasts are not far from the average revision made by Statistics Finland, which is an encouraging result in light of actual implementability of the method. Finally, we find that using the factor nowcasts of TIO in the computation of quarterly GDP allows us to reduce the publication lag without loss of accuracy, compared to the current flash estimates of GDP.

The remainder of the paper is structured as follows: in Section2 we present the two-stage statistical model employed to construct nowcasts of TIO: In Section 3 we describe the data and, in particular, how 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 employed nowcasting model consists of two stages. In the first one, we extract common factors from a large dataset of firm-level turnovers (Section 2.1). When the factors are extracted, they are used in nowcasting regressions (Section 2.2) to construct nowcasts of the variable of interest, which in this study is the monthly Finnish economic activity.

2.1 Factor Extraction

The factors are computed as in the factor model of Stock and Watson (2002a). There are multiple reasons for this choice. The datasets we use to compute the TIO estimates are very large. The sale inquire includes almost 2000 firms, hence we need a model which can handle such large cross sections. While the model of Banbura and Modugno (2010) can also handle various data problems and it is used widely in the nowcasting literature, it is too computationally demanding for this application. In the Stock and Watson (2002a), a large dataset follows a factor model with r latent factors included in F_t. Defining now X_t as the dataset containing N series of firm-level turnovers, we can write the factor model as follows

$$X_t = \Lambda F_t + e_t, \qquad (1)$$

where $\boldsymbol{\Lambda}$ is the matrix of factor loadings and e_t is the N×1 vector of idiosyncratic components. The idiosyncratic components are allowed to be both serially and cross-sectionally correlated, making this model resembling the approximate factor model by Chamberlain and Rothschild (1983). The factors are estimated by principal components, i.e. $\hat{F}_t$ is given by the eigenvectors corresponding to the largest eigenvalues of the $T \times T$ matrix XX'. This is a computationally handy procedure, because we do not have to deal with very big matrices in the estimation, despite the very large cross section of firms.

A common feature of datasets used in nowcasting exercises, similar to this paper, is the presence of jagged edges and missing values. The basic principal component estimation requires balanced dataset (i.e. all the series should be of the same length). We deal with the missing values problem in two ways in this study. In the first method we simply create a balanced dataset by taking a subset of variables from the original data. In this way we do not have to perform missing value imputation, with the associated estimation errors and computational intensity, but we have to give up a large part of the original data, at least for the very early estimates. We refer this methodology as a balanced method later on.

As alternative procedure, we use the regularized iterative principal component analysis (PCA) algorithm (see details in Josse and Husson, 2012b). This method is preferred to the simple EM iterative PCA presented in Stock and Watson (2002b), because it is targeted for datasets with many missing values, which is the case in the data to be analyzed in Section 3 and 4. Moreover the regularized iterative PCA method performs better regarding the overfitting problem.

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 whereas Stock and Watson (2002b) suggest to use an initial balanced dataset to compute the first estimate of the factors. In the second step, we use the estimated factors to impute the missing data following the equation:

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

where $\hat{X}_{tk}$ is a missing value at time t for 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 imputed values. We iterate these three steps until we reach convergence.

The basic intuition of the regularized PCA algorithm is that if there is a lot of noise in the data, or equivalently the structure of the data is too weak (for example lots of missing values), the algorithm weights less the principal component imputation (the $\sum_{s=1}^{S} \hat{F}_{ts} \widehat{\mathbf{\Lambda}}_{ks}$ in equation (2)) and it tends to impute the simple mean of the variable ($\hat{\mu}_k$). If the noise in the data is small, then this algorithm converges to the simple EM algorithm of Stock and Watson (2002b). More formally, the regularized PCA algorithm shrinks the principal component part of the imputation step getting

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

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

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

$$y_t = \beta_v \hat{F}_{t+v} + \epsilon_{t+v} \qquad (4)$$

where y_t measures economic activity, with t being the reference moth we are interested in, ϵ_{t+v} is the nowcasting error and v is the period in which we compute our nowcasts (i.e. how many days after the end of the reference period we compute the estimate). In our application, we estimate equation (4) nine times for each period, that is at t+5, +10, +15 up to t+45 days after the end of the reference moth (see details in Section 3). We do not compute factor estimates after t+45 because by that time economic activity indicators are usually released. The mean squared error minimizing nowcasts are constructed as $\hat{y}_{t|\ v} = \hat{\beta}_v \hat{F}_{t|\ v}$, where $\hat{y}_{t|\ v}$ denotes the predicted value at time v and the parameters β are estimated by the Ordinary Least Squares (OLS).

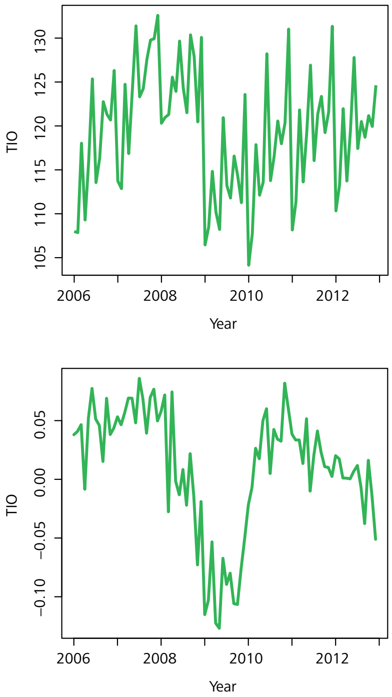
One important issue in the estimation process stems from the factor selection, i.e. how many factors should be included in $F_{t|\ v}$. For the balanced method, the factor selection can be based on information criteria, as the Bayesian Information Criteria (BIC) or factor based regression criteria suggested by Groen and Kapetanios (2009). We also compute nowcasts based on 10 factors, and check the out-of-sample performance of the various models. The estimation of the number of factors is even more delicate matter when we deal with missing values replacement (see Section 2.1). Josse and Husson (2012a) provides an algorithm that estimates the optimal number of principal components for a given dataset presenting missing values.

3 Data Description

The variable we are interested in this study is the Trend Indicator of Output (TIO), both in levels and year-on-year growth rates, measuring Finnish economic activity in a monthly basis. The sample period stars in January 1998 and ends in December 2012. In the out-of-sample nowcasting experiment in Section 4, we estimate TIO starting from January 2006, giving us a total of 84 periods to nowcast.

The TIO is currently released by Statistics Finland with two different schedules, based 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, the figure is published 45 days after the end of the reference period[3]. 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 data and value data which are aggregated to a 2-digit level. Primary sources of data for private manufacturing are the preliminary turnover indices (which are accumulated quickly), while for the public sector the main sources are preliminary wages and salaries. In the Finnish system of national accounts, flash quarterly GDP estimate for Finland is published 45 days after the end of the reference quarter and it is based on the TIO figures. Below, in Figure 3.1, we depict the time series of TIO, both in level and percentage changes during the nowcasting period.

Figure 3.1 Plots of the TIO during



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

It is interesting to notice the big drop in economic activity during the recent recession and the fact that the level value is still somewhat below the pre-recession period.

A major contribution in this study is to use firm-level data in factor estimation for nowcasting. Due to their timeliness, the firm-level turnover data appears as an interesting alternative to the previously considered datasets used in factor extraction (see Giannone et al. (2008)). This data is accumulated right after the end of the reference month and the date on which a firm sends its data to Statistics Finland is well documented and collected in a dataset. Thanks to these reports, we can replicate closely a real-time environment, and together with the high quality of this data, they motivate us to consider Finnish data throughout this paper.

In our nowcasting experiment we simulate the data accumulation process by creating different real-time datasets of turnover indeces available at different periods. For each month we create nine (i.e. nine different v in (4)) different datasets corresponding to turnover indices available at t+5, +10 and so on. For example, when we estimate TIO in December 2009 at t+20, we base our estimation on turnovers available by Januay 20th and we use only turnovers of private firms as the explanatory dataset.

While it is true that other data could be useful, we want to extract and isolate as much as possible the ability of this firm-level data to give early signals of the TIO. Given the novelty of this dataset in a nowcasting application, it is useful to check for its predictive power in the most straightforward way and adding more predictive variables would complicate the analysis. Moreover, focusing on turnovers indices allows us to have a very precise replication of the data accumulation, which becomes much more cumbersome when some additional data sources are also examined. This complication is avoided in a real time application, in which we do not need to replicate the data accumulation process.

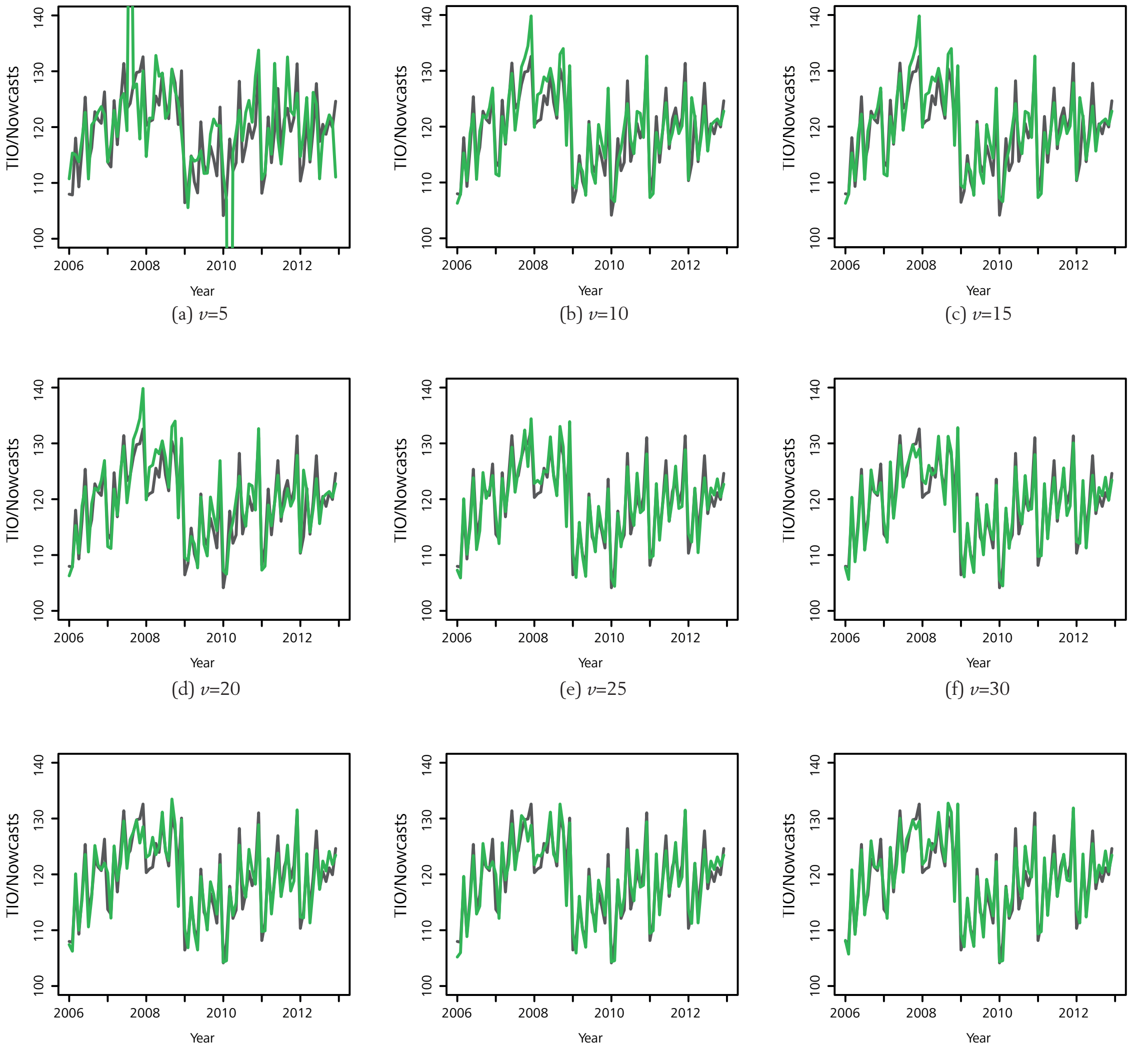
The original turnovers dataset contains more than 2000 firms, but many of these series present extremely high number of missing values. Because we want to compute the nowcasts starting from the beginning of 2006 and start the estimation period as early as possible, we exclude many firms from the dataset. We keep firms that started reporting already in 1998 and reported at least up to the end of 2005. This gives us an initial balanced dataset for the estimation. The remaining dataset includes 579 firms. The volume of turnovers of these firms amounts to 45% of total turnovers, in the beginning of the dataset, increasing up to 64% at the end of 2005 and up to 77% of total turnovers at the end of 2012. While the loss of information seems quite large in the beginning of the sample, the later periods of the dataset seem to contain a large fraction of the total turnovers of the original dataset. In the data appendix we report additional information about the data accumulation process, e.g. the percentage of firms reporting by v days after the end of the reference period on average, and the plot of the cumulative eigenvalues for the turnovers dataset. These statistics are useful to analyze how much the data accumulation affects our estimates and how much the information contained in this large disaggregated datasets can be squeezed into few factors.

4 Empirical Results

We compute nowcasts by following the methods described in Section 2. The starting in-sample period goes from January 1998 to December 2005, whereas the nowcasts start from January 2006 and we re-estimate the model forward, using an expanding window, up to December 2012. We start analyzing the empirical results by having a look at the plots of the nowcasts against the original series. While this is an informal method to analyze the results, visualizing nowcasts can give a lot of insights on their performance. In Figure 4.1 we show the nowcasts for the TIO in levels for the model using the factors selected by BIC. In this section we only report the nowcasts obtained by using the BIC. In this section we only report the nowcasts obtained by using the BIC criterion, because using the Groen and Kapetanios (2009) criterion led to the same results. We compare the prediction performance based on the root mean squared forecast error (RMSFE) using an AR(p) model and a random walk (RW) as benchmark, where p in the AR model is selected through BIC. Moreover, we compute the mean absolute percentage error of predictions, to shed some light on the actual applicability of the method.

In red we have the nowcasts performed with the balanced method and with factors selected with the BIC. We immediately see that at t+5, i.e. after 5 days from the end of the reference period, the nowcasts are pretty inaccurate. In particular around the half of 2007 and the beginning of 2010 we have two extreme nowcasts. If we omit those two periods, it seems that the nowcasts are able to detect the overall trend of the series, but they are still in general inaccurate. Remember that in the case v=5, the nowcasts are based on a very small set of firms turnovers. Already at t+10 and t+15, we have a fairly large improvement. There seems to be much less implausible spikes and the nowcasts seem to track much better the original series. In both cases, there is still a pretty large spike around the end of 2007 prediction but it seems to disappear from the nowcasts done after 20 days or more. Another interesting feature is that there are no visible improvements by going over 20 days after the end of the reference period. This indicates that the t+20 selection might be optimal for the factor model in terms of the tradeoff between timeliness and the accuracy of nowcasts. This selection is able to pick up the most interesting co-movements in the turnover dataset and it also appears that to increase the accuracy of the nowcasts even more, we might need to augment the model with some additional predictive variables.

Figure 4.1 Nowcasts computed with the balanced method at $t + v$



(a) v=5

(b) v=10

(c) v=15

(d) v=20

(e) v=25

(f) v=30

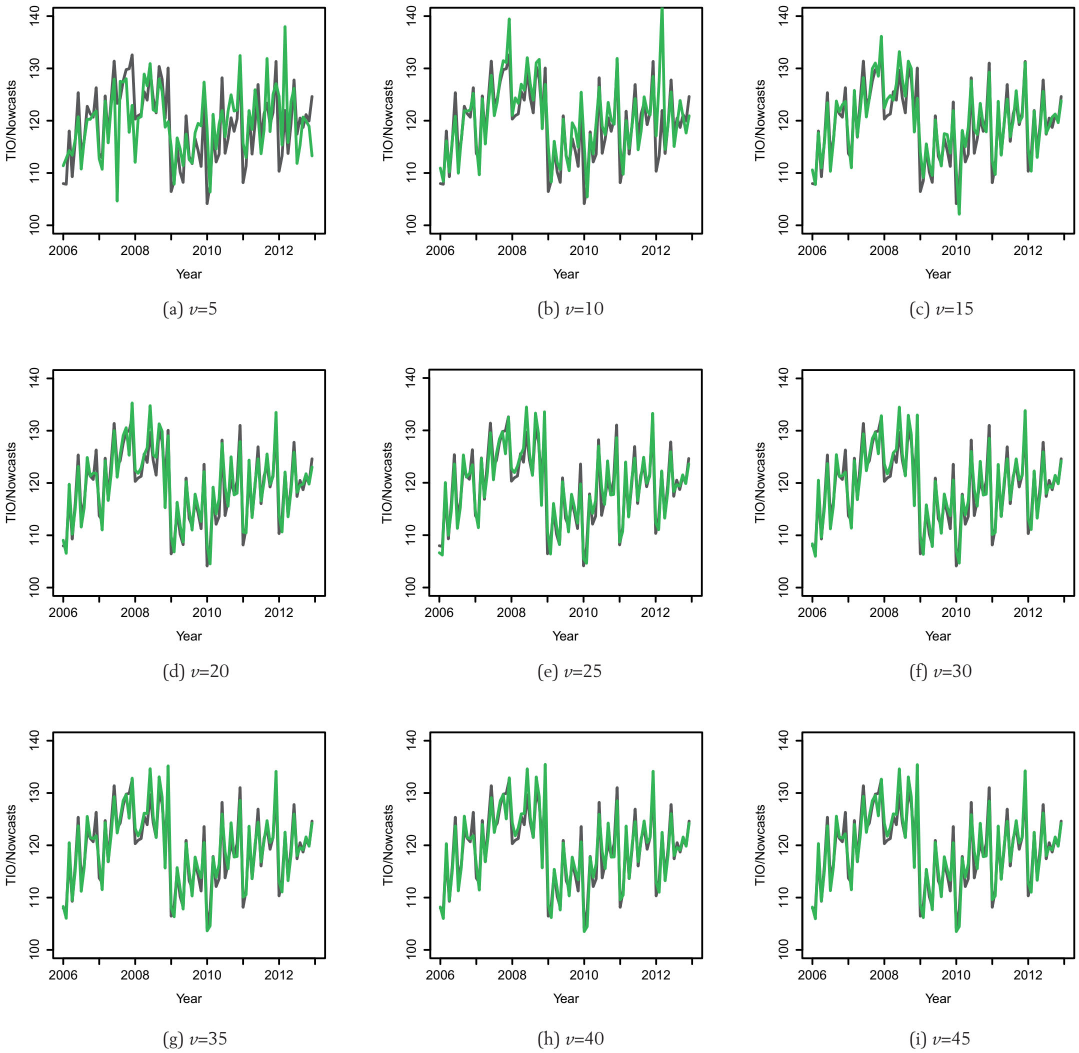
(g) v=35

(h) v=40

(i) v=45

Next we report the plots for nowcasts based again on the factors extracted using the regularized EM algorithm. As above, the factors are again selected using the BIC criterion. In this way we perform missing value imputation, with the related prediction error, but we can use larger datasets even at earlier times.

Figure 4.2 Nowcasts computed with the EM method at $t|\ v$



(a) v=5 (b) v=10 (c) v=15

(d) v=20 (e) v=25 (f) v=30

(g) v=35 (h) v=40 (i) v=45

We immediately see that there is an improvement in nowcasts computed at t+5, even though, similarly as in Figure 4.1, they remain inaccurate. The nowcasts based on the regularized iterative PCA seem to perform well, being able to predict also the very end of the sample, which is something that the balanced method seems to have difficulties with.

Even though plots can give a general impression how the methods perform, we still need to use some numerical evaluation criteria to judge the out-of-sample performance of the models at hand. We use two different measures: the root mean square forecasts error, and the percentage absolute error. The first one is defined as

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

In the subsequent tables, we report the relative RSMFEs which are computed relative to the RMSFE of two benchmark models, the AR(p) model and the random walk. Thus, a value below 1 indicates that our nowcasting model gives better predictions compared to the benchmark models. The other measure, the mean absolute percentage error (given by MAPE=$\sqrt{\frac{1}{T}\sum_{t=1}^{T}|\,(y_t-\hat{y}_{t|\,v})/y_t|\,}$), gives an idea of how far are our estimates, on average, from the true value. It gives a good indication of the method performance in light of a practical implementation. Moreover, we rely on the Diebold and Mariano (1995) test when comparing predictive accuracy of two non-nested nowcasting models (we us the AR(p) model as alternative).

In Table 4.1 we report the relative RMSFE for the balanced and EM method using 10 factors and the BIC selected factors.

Table 4.1:
Relative (AR) RMSFE for nowcasts of TIO in levels. In the table *,**,*** indicate rejection of equal predictive ability between AR model and factor model at 10, 5 and 1 % statistical significance level respectively

v	BIC factor (Bal.)	10 factors (Bal.)	BIC factors (EM)	10 factors (EM)
5	3.75	3.75	1.36	1.22
10	0.74**	0.76**	1.008	0.77**
15	0.56***	0.70***	0.61***	0.64***
20	0.55***	0.60***	0.53***	0.61***
25	0.51***	0.61***	0.49***	0.64***
30	0.51***	0.60***	0.50***	0.64***
35	0.57***	0.59***	0.50***	0.64***
40	0.55***	0.60***	0.51***	0.64***
45	0.51***	0.60***	0.50***	0.63***

This table gives us already few insights. First of all, it seems that the methods proposed here are able to beat the benchmark (AR(p) model) for most of vs. This is reflected in the fact that the relative RMSFEs are consistently below unity. Only the nowcasts performed five days 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 a moderate advantage over the basic method. Another interesting aspect is that the predictive performance does not improve much after t+20. For our nowcasting application, the principal components are able to estimate the important underlying factors just by using a subset of firms, without needing the complete dataset. The overall better predictive performance is also confirmed by the Diebold-Mariano test. This result applies also when a random walk with drift is used as a benchmark instead of the AR(p) model (see Table 4.2).

Table 4.2:
Relative (RW) RMSFE for nowcasts of TIO in levels. In the table *,**,*** indicate rejection of equal predictive ability between AR model and factor model at 10, 5 and 1 % statistical significance level respectively

v	BIC factor (Bal.)	10 factors (Bal.)	BIC factors (EM)	10 factors (EM)
5	1.81	1.81	0.66***	0.59***
10	0.36***	0.36***	0.48***	0.37***
15	0.27***	0.34***	0.29***	0.31***
20	0.26***	0.29***	0.25***	0.29***
25	0.25***	0.29***	0.23***	0.31***
30	0.24***	0.29***	0.24***	0.30***
35	0.27***	0.28***	0.24***	0.30***
40	0.26***	0.29***	0.24***	0.30***
45	0.24***	0.29***	0.24***	0.30***

Now let us analyze the nowcasting performance of the model for the monthly year-on-year changes of TIO. The models used are in the same line as the ones used so far. Below we report the relative RMSFE table for the balanced and EM models against the AR benchmark.

Table 4.3:
Relative (RW) RMSFE for nowcasts of TIO in percentage changes. In the table *,**,*** indicate rejection of equal predictive ability between RW model and factor model at 10, 5 and 1 % statistical significance level respectively.

v	BIC factor (Bal.)	10 factors (Bal.)	BIC factors (EM)	10 factors (EM)
5	1.05	1.81	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***

Again the methods considered here beat the AR(p) benchmark except for t+5 estimates. Notice that the EM estimates for 10 factors now also beat the balanced method nowcasts. Furthermore, nowcasts based on a richer model (with factors selected by the BIC) perform better than the more parsimonious model in later period (this is valid for the EM case). Overall the relative RMSFE follows the same patter as in Table 1. In other words, the models studied here beat the benchmarks except for the nowcasts at t+5, confirming the fact that at t+5 there is not enough firm-level data accumulated to estimate the underlying factors and hence nowcast the TIO accurately.

It is also important to have an idea of how much our predictions deviate from the actual (revised) values of TIO, to evaluate how well the models would perform in practice. Tables 4 and 5 report the mean absolute percentage errors for the TIO level and year-on-year percentage changes.

Table 4.4:
Mean absolute percentage errors for nowcasts of TIO in level

v	BIC factor (Bal.)	10 factors (Bal.)	BIC factors (EM)	10 factors (EM)
5	0.04	0.04	0.03	0.03
10	0.019	0.020	0.020	0.021
15	0.015	0.018	0.016	0.017
20	0.015	0.016	0.014	0.016
25	0.014	0.016	0.013	0.017
30	0.014	0.015	0.013	0.017
35	0.015	0.015	0.013	0.017
40	0.015	0.015	0.013	0.016
45	0.014	0.015	0.013	0.016

The EM based predictions seem to perform better at t+5 and also in later periods than the balanced method. Also, more parsimonious models seem to create worse estimates than models with more factors included in the nowcasting model (4).

Table 4.5
Mean absolute percentage errors for nowcasts of TIO in percentage changes

v	BIC factor (Bal.)	10 factors (Bal.)	BIC factors (EM)	10 factors (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	0.014	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	0.013

What do we gather from these tables? The usual percentage deviation from the actual revised TIO value is 1.4%. Usual revision done by Statistics Finland is around 0.9%, so our estimates do somewhat worse than the ones made by Statistics Finland. This is expected, because the Statistics Finland revisions are based on the actual figures, which have substantial lags in the publication 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, so considering the reduction in the publication lag, this method provides an attractive alternative for the existing method currently used.

So far we focused on the nowcasts of monthly TIO, but a very interesting application of this methodology lies in the prediction of quarterly GDP. In particular we can use the nowcasts of TIO to compute early estimates of GDP, with shorter publication lags. The current flash estimate of GDP computed by Statistics Finland is published around 45 days after the end of the reference quarter. With the method presented in this paper, we can shorter considerably this publication lag. 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. Below we report the table with 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 we use is based on the EM method and we report results based on t+25 estimates. The number of factors is selected with BIC.

Table 4.6
Mean absolute percentage error for nowcasts and flash estimates of quarterly GDP in year-on-year percentage changes

v	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

The results presented in the previous table are very encouraging. It seems that we can shorten considerably the publication lag without a major increase in the estimation error. In particular, in the six years between 2006 and 2012 the measurement error is substantially equal between the factor model estimates and the current flash estimates, while the factor method manages to beat the current estimates for the year 2012. Based on these results, the factor model nowcasts provide a competitive method for quarterly GDP estimation. Of course there is a lot of space for improvements, e.g. new data can be included in the factor estimation.

5 Conclusions

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

We find that these two methods beat the benchmark models for all estimation periods except for very early periods close to the end of the month we want to nowcast, We also find that the EM method does provide better nowcasts compared to 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. In particular we see that the nowcasts computed with this method allow a substantially shorter publication lag (in our case 20 days reduction in the publication). The main finding of this study is that the factors extracted from the large micro dataset are useful when 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. Together with the firm- level turnovers, we could also include macroeconomic and financial variables in our nowcasting model. Moreover, we might use the factors and the TIO estimates obtained in this exercise in a wider nowcasting application. Very early nowcasts can be based on surveys and financial variables, but as time goes on we can add the TIO estimates as indicators in the nowcasting models. Also the regression used for the prediction can be modified, for example by adding lags of the dependent variable or even of the factors. We could also use models which take into account factor dynamics in the factor estimation. For example the two steps estimator proposed by Doz et al. (2011) could be used, but computational complexity has to be considered.

References

A, K.N &Trovik, T.J (2008). Estimating the output gap in real time: A factor model approach. Working Paper 2008/23.

Alessi, L. & Barigozzi, M. & Capasso, M. (2006). A dynamic factor analysis of business cycle on firm-level data. LEM Working Paper Series 27, Laboratory of Economics and Management Sant'Anna School of Advanced Studies, Pisa, June 2006.

Altissimo, F. & Cristadoro, R. & Forni, M. & Lippi, M. & Veronese, G. (2007). New eurocoin: Tracking economic growth in real time. Temi di discussione (Economic working papers) 631, Bank of Italy, Economic Research and International Relations Area,.

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

Banbura M. & Modugno, M.(2010). Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data. Working Paper Series 1189, European Central Bank, May 2010.

Banbura, M. & Giannone, D. & Reichlin, L. (2010). Nowcasting. Working Papers ECARES 2010-021, ULB – Universite Libre de Bruxelles.

Chamberlain, G. & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica, 51(5): 1281–304, September 1983.

Diebold, F.X & Mariano, R.S (1995). Comparing predictive accuracy. Journal of Business and Economics Statistics, 13(3), July 1995.

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, September 2011.

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

Giannone, D. & Reichlin, L & Small, D. (2008). Nowcasting: the real time informational content of macroeconomic data releases. ULB Institutional Repository 2013/6409, ULB – Universite Libre de Bruxelles.

Groen, J.J.J & Kapetanios, G.(2009). Model selection criteria for factor-augmented regressions. Staff Reports 363, Federal Reserve Bank of New York.

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

Josse, J. & Husson, F. (2012b). Handling missing values in exploratory multivariate data analysis methods. Journal de la SFdS, 153(2): 79–99, 2012.

Martinsen, K. & Ravazzolo, F. & Wulfsberg, F. (2011). Forecasting macroeconomic variables using disaggreagate survey data. Working Paper 2011/4, Norges Bank, April 2011.

Matheson, T. & Mitchell, J. & Silverstone, B. (2007). Nowcasting and predicting data revisions in real time using qualitative pane survey data. Reserve Bank of New Zealand Discussion Paper Series DP2007/02, Reserve Bank of New Zealand, Feb 2007.

Mitchell, J. & Weale, M. (2005). Quantitative inference from qualitative business survey panel data: a microeconometric approach. NIESR Discussion Papers 261, National Institute of Economic and Social Research, September 2005.

Modugno, M. (2011). Nowcasting inflation using high frequency data. Working Paper Series 1324, European Central Bank, April 2011.

Proietti, T. (2008). Estimation of common factors under cross-sectional and temporal aggregation constraints: Nowcasting monthly gdp and its main components. MPRA Paper 6860, University Library of Munich, Germany, January 2008.

Stock, J.H & Watson, M.W (1989). New indexes of coincident and leading economic indicators. In NBER Macroeconomics Annual 1989, Volume 4. NBER Chapters, pages 351–409. National Bureau of Economic Research, Inc, December 1989.

Stock, J.H & Watson, M.W (2002a). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics, 20(2): 147–62, April 2002.

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, December 2002.

Appendix 1

One of the nice features of the data used in this analysis lays in the possibility of tracking the data accumulation faced by Statistics Finland. It is interesting how the data accumulation evolves over time, reflecting the dynamics of the information available to the data producer. Below we report the table with the average number and the percentage of firms sending their turnovers data to Statistics Finland at different time points after the end of a given period. We also include the percentage of total turnovers reported by a given date, to check whether there is some relation between the size of firms and the timeliness of their reports

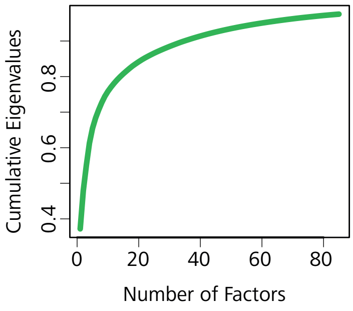
Table A.1
Accumulation of turnovers data by v days after the end of the reference month

v	5	10	15	20	25	30	35	40	45
Number of Firms	35	125	262	389	432	454	460	465	468
Percentage Firms	0.07	0.26	0.56	0.83	0.91	0.96	0.97	0.98	1.00
Percentage Turnovers	0.07	0.25	0.57	0.84	0.92	0.97	0.97	0.98	1.00

The accumulation of the data seems to become very slow after v=20 or v=25 days after the end of the reference month. This reflects the fact that the nowcasting performance does not improve much after v=20. Moreover the percentage of firms reporting and the percentage of turnovers accumulated are very close to each other. This indicates that there is not a specific pattern of which kind of firms send their turnovers first. If the largest firms would send their turnovers first, then we would find that the turnover accumulation would be faster than the percentage of firms reporting.

Another interesting question related to this highly disaggregated dataset is how much information can be squeezed into fewer variables, the factors. To shed some light on this matter, we report below the plot of cumulative eigenvalues for the turnovers dataset in December 2012 of firms reporting by January 31st, so the last, and most extensive vintage available.

Figure A.1 Cumulative eigenvalues plot



This plot gives us a rough idea of how much variance in the turnover dataset is explained by the common factors. It seems that after hitting 80% of explained variance (see also Table 7), around 20 factors, the cumulative eigenvalue curve becomes rather flat. This suggests that we cannot rely only on very few factors in our nowcasting model. However, we find that a rich model, with more than 20 factors in the prediction regression, performs well, so it seems that we do not encounter overfitting issues.

Table A.2
Percentage of variance in the turnovers dataset explained by common factors

Number of factors	1	2	3	4	5	6	7	8	9	10
Explained variance	0.37	0.47	0.55	0.61	0.65	0.68	0.7	0.72	0.74	0.75

Number of factors	11	12	13	14	15	16	17	18	19	20
Explained variance	0.77	0.78	0.79	0.79	0.8	0.81	0.82	0.82	0.83	0.84

Tilastokeskus
Statistikcentralen
Statistics Finland