ETLA Working Papers

No. 46

10 January 2017

Paolo Fornaro
Henri Luomaranta
Lauri Saarinen

Nowcasting Finnish Turnover Indexes Using Firm-Level Data

Suggested citation: Fornaro, Paolo, Luomaranta, Henri & Saarinen, Lauri (10.1.2017). "Nowcasting Finnish Turnover Indexes Using Firm-Level Data". ETLA Working Papers No 46. http://pub.etla.fi/ETLA-Working-Papers-46.pdf

ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY

ISSN-L 2323–2420
ISSN 2323–2420 (print)
ISSN 2323–2439 (online)

Nowcasting Finnish Turnover Indexes Using Firm-Level Data*

Paolo Fornaro[a], Henri Luomaranta[b], Lauri Saarinen[c]

[a]Research Institute of the Finnish Economy

[b]Statistics Finland and University Of Toulouse 1

[c]Statistics Finland

January 2017

Abstract

We adopt a series of shrinkage and factor analytic methodologies to compute nowcasts of the main Finnish turnover indexes, using continuously accumulating firm-level data. We show that the estimates based on large dimensional models provide an accurate and timelier alternative to the ones produced currently by Statistics Finland, even after taking into account data revisions. In particular, we find that the turnovers for the service sector can be estimated with high accuracy five days after the reference month has ended, giving more accurate and faster predictions compared to the first official internal release. For other sectors, the large dimensional models provide a good nowcasting performance, even though there is a timeliness-accuracy trade off. Finally, we propose a factor-based methodology to improve the accuracy of the current flash estimates by imputing part of the data sources, and find that we are able to provide better predictions in a more expedited fashion for all sectors of interest.

JEL Classifcation Code: C31, C53, C55

Keywords: dynamic factor models; firm-level data; nowcasting; shrinkage

1 Introduction

The vast literature on nowcasting (see, among many others, Evans, 2005, Altissimo, Cristadoro, Forni, Lippi, and Veronese, 2010, Giannone, Reichlin, and Small, 2008, Aruoba, Diebold, and Scotti, 2009) has focused on solving the issue of timeliness of official data releases. Throughout the years, it has made some very useful advancements with the availability of high speed computing and the development of methods that are capable of dealing with high dimensional econometric problems. It is a major concern for

*We want to thank the Eurostat Big Data ESSNet for providing financial support to Henri Luomaranta and Lauri Saarinen.

informed policy making that many important indicators are published either with long delay, or are inaccurate and consequently revised multiple times. Currently, Statistics Finland publishes turnover indexes with a lag of 75 days, although the first internal estimates are produced 45 days after the end of the reference month. However, Eurostat requests the publication lag for these indicators to be considerably shortened (to 60 or 45 days, depending on the series) and recommends a fairly low revision error (around 1% point absolute deviation on average). Many internal and external users of these indexes would greatly benefit from these indicators being produced more quickly and more accurately. Turnover indexes are widely followed in their own right, but are also used as source material for producing the Trend Indicator of Output (TIO, i.e. the Finnish monthly economic activity indicator), and various volume indexes (e.g. the industrial production indicator or volume indexes for the trade and construction sectors). Hence a faster availability of turnover indicators would allow Statistics Finland to accelerate the production of many important indexes, such as the TIO and consequently the GDP.

We investigate turnover indexes in services, manufacturing, wholesale and retail trade, and construction[1], and demonstrate that high dimensional models using firm-level data are able to beat the benchmark ARIMA models at $t|5$ days (i.e. 5 days after the end of the reference month), and achieve a relatively high level of accuracy at $t|10$. We consider various shrinkage models (ridge regression, the lasso and the elastic-net) and the factor model of, e.g., Stock and Watson (2002b). One of the most notable findings is that we are able to nowcast the services turnover index with very high precision already at $t|10$, offering a very useful, almost real time, alternative to the current publications. Furthermore, we propose a methodology, based on factor models, by which the statistical institute could improve the accuracy and the timeliness of their first internal estimate by imputing part of the source data.

The rest of the paper is structured as follows: in Section 2 we present the models that we adopt to create the estimates of the turnover indexes. Section 3 contains a short description of the data and of our empirical application. In Section 4, we report the results of the nowcasting exercise and Section 5 concludes.

[1] The indexes examined are services (HIJLMNRS), manufacturing (C), wholesale and retail trade (G), and construction (F) in the NACE rev.2 nomenclature.

2 Methodology

In this section, we briefly describe the shrinkage techniques that we adopt in the nowcasting exercise. The use of various shrinkage methodologies in order to obviate the curse of dimensionality, due to a large number of model parameters to be estimated, has been the focus of a wide literature. We use the well known ridge regression, the least absolute shrinkage and selection operator (lasso, Tibshirani, 1996) and the elastic-net of Zou and Hastie (2005). Additionally, we extract the common factors underlying our dataset following Stock and Watson (2002a).

We report below the shrinkage methods used in our analysis.

Ridge Regression

The basic idea of the ridge regression methodology is to penalize the size of the regression coefficients and shrink them toward 0. In practice this is obtained by minimizing

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{K} \beta_j^2, \tag{1}$$

where $\mathbf{y}$ is the variable we want to predict and $\mathbf{X}$ is the matrix of K predictors. λ determines the degree of shrinkage (i.e. how much we are forcing the parameters to be near 0). In a Bayesian framework this can be interpreted as imposing a prior following a normal distribution with mean 0 and variance proportional to λ. The solution of the minimization problem of gives us:

$$\widehat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

where $\mathbf{I}$ is $K \times K$ identity matrix. Notice that the ridge regression does not attempt to isolate the variables with good predictive power, instead it is aimed at regularizing the large dimensional regression solution.

Lasso

This shrinkage estimator was introduced in Tibshirani (1996). The main idea of the methodology is to produce models where the parameters of irrelevant variables are estimated to be exactly zero, leading to a variable selection setting. The minimization

problem behind the lasso can be specified as

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{K} |\beta_j|. \tag{2}$$

Even though lasso has many benefits, it does have some drawbacks. For example if there are many multicollinear predictors, lasso estimation will lead to select only one of these useful predictors, disregarding all others. The elastic-net of Zou and Hastie (2005) is helpful in this scenario.

Elastic-Net

Introduced in Zou and Hastie (2005), the elastic net combines ridge-regression and the lasso. It is based on the following minimization problem

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda_1 \sum_{j=1}^{K} |\beta_j| + \lambda_2 \sum_{j=1}^{K} \beta_j^2 \tag{3}$$

One of the main benefits of the elastic-net is that it is better suited in a scenario where the predictors are strongly correlated and has been shown to work better when the number of predictors is larger than the number of observations. Given that our firm-level data is based on turnovers, we expect their year-on-year growth rates to be fairly cross-correlated, due to the impact of aggregate business conditions. Moreover, especially when looking at firm data accumulated many days after the end of the reference month, we expect the number of firms in our predictors set to be larger than the number of time series observations.

All models are estimated using the 'glmnet' package for R. The details of the computation algorithm are given in Friedman, Hastie, and Tibshirani (2010). The degree of shrinkage (i.e. the values of λ, λ_1 and λ_2 in (1)-(3)) is selected through 10-fold cross validation.

In addition to the various shrinkage methodologies, we adopt the factor model of Stock and Watson (2002a) in our nowcasting exercise. In particular, we use principal components estimation to extract the common factors underlying our data. We select the number of factors, using the criteria proposed in Bai and Ng (2002), at each estimation round.

3 Data and empirical application

The data used in this analysis originates from the sales inquiry carried out by Statistics Finland. This dataset covers around 2,000 of the most important enterprises in their respective industries, representing ca. 70% of total turnovers. The data is available soon after the end of the month of interest and a considerable share of the final data is accumulated around 20 days after the end of the reference month. When firms send their data to Statistics Finland, a log of the day and time of the report is recorded. In this fashion, we can simulate realistically the data accumulation faced by Statistics Finland. A similar dataset is adopted in Fornaro (2016), even though the focus in that work is the use of common factors extracted from the firm-level data to nowcast the Finnish monthly economic activity indicator. We require that firms have long time series (starting in 2000), and that they have reported sales figures by the date we extract their information from the database. This leads us to have on average 70 firms at $t|5$, 268 firms at $t|10$, 715 firms at $t|20$ and 799 firms at $t|26$ in the data.

We use these monthly firm-level data on sales, analyzed at the premises of the statistical office after anonymization, to estimate the year-on-year monthly growth rate of the four turnover indicators considered. We create four estimates for all the months in the period January 2013 - July 2016 . The first one at $t|5$, the second at $t|10$, then at $t|20$ and finally at $t|26$, in order to compare the models when dealing with different degrees of informational content. To evaluate the predicting performance of our models we rely on the mean squared forecast error and the mean absolute error:

$$MSFE = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_{t|\nu})^2 \tag{4}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|(y_t - \hat{y}_{t|\nu})| \tag{5}$$

Notice that $\hat{y}$ and y indicate our nowcast of the turnover index growth rates and their actual value (the latest revision), respectively. As benchmark, we use an ARIMA model with specification selected automatically for each nowcast period. The ARIMA model is based solely on the past values of the aggregate turnover indexes.

4 Empirical Results

In Table 1, we report the $MSFE$s relative to the ARIMA benchmark. A value less then one indicates a superior performance of the large dimensional models.

	$t\|5$	$t\|10$	$t\|20$	$t\|26$
		Manufacturing		
Elastic-Net	0.479	0.495	0.354	0.346
Lasso	0.526	0.579	0.422	0.399
Ridge	0.421	0.353	0.326	0.327
Factors	0.913	0.604	0.471	0.480
		Services		
Elastic-Net	0.906	0.831	0.390	0.395
Lasso	0.931	0.857	0.388	0.380
Ridge	0.756	0.548	0.475	0.479
Factors	1.343	0.900	0.588	0.598
		Trade		
Elastic-Net	0.708	0.567	0.298	0.293
Lasso	0.735	0.596	0.306	0.289
Ridge	0.517	0.536	0.390	0.365
Factors	0.805	0.499	0.376	0.351
		Construction		
Elastic-Net	0.875	0.931	0.697	0.796
Lasso	0.869	0.945	0.720	0.812
Ridge	0.721	0.805	0.687	0.681
Factors	0.942	0.698	0.504	0.501

Table 1: Relative $MSFE$s of the large dimensional models against the ARIMA benchmark. Values below one indicate a better performance of the firm-level data based model.

As we can see from these results, all models based on firm-level data outperform the ARIMA benchmarks at all times of estimation. The only exception is in the service sector for the estimates obtained by the factor model 5 days after the end of the reference month. Moreover, it is interesting to notice that there is a moderate drop in the MSFE between $t|5$ and $t|10$, for most methods and sectors, but there is a much more evident shift at $t|20$. Going from $t|20$ to $t|26$ does not seem to be beneficial in terms of improving the nowcasting accuracy. This indicates that additional firm-level information is not helpful in improving the predictions of the turnover indexes, once we include a certain number of firms in the estimation. In particular, it seems that the models are able to extract the crucial firm-level information already at $t|20$.

Even though the results of Table 1 highlight the advantages of using fast accumulating firm-level data in nowcasting aggregate indexes, we need to be cautious toward their

practical significance. While they consistently outperform the estimates that we would obtain by using the aggregate turnovers indexes directly, we need to still check if they deviate too much from the actual values of the target indicators. To do this, we compute the mean absolute error, giving us the percentage point difference between the predictions and the actual values of turnover indexes year-on-year growth rates.

Importantly, we do a pseudo-real time analysis. In particular, we use realistic vintages of the data, reflecting the information available to the statistical institute at the time that the nowcast would be computed. The vintages are available from January 2013 for the trade sector and from February 2013 for the services, manufacturing and construction, hence we compare the nowcasting performance using these restricted time windows. Results are reported in Table 2.

	$t\|5$	$t\|10$	$t\|20$	$t\|26$
		Manufacturing		
Elastic-Net	2.097	2.324	2.017	1.999
Lasso	2.268	2.369	2.177	2.122
Ridge	**1.833**	**1.861**	**1.712**	**1.675**
Factors	2.739	2.582	2.166	2.036
t|45 (current method)	**1.084**			
		Services		
Elastic-Net	1.471	1.248	**0.979**	**0.984**
Lasso	1.508	1.251	0.980	0.999
Ridge	**1.269**	**1.212**	1.071	1.076
Factors	1.707	1.348	1.183	1.166
t|45 (current method)	**1.730**			
		Trade		
Elastic-Net	1.978	1.697	**1.165**	**1.162**
Lasso	2.045	1.758	1.172	1.175
Ridge	**1.762**	**1.650**	1.292	1.297
Factors	2.272	1.726	1.441	1.443
t|45 (current method)	**0.895**			
		Construction		
Elastic-Net	3.302	3.620	2.851	3.054
Lasso	3.358	3.641	2.800	3.098
Ridge	**2.982**	3.308	3.029	2.988
Factors	3.466	**3.215**	**2.487**	**2.359**
t|45 (current method)	**2.555**			

Table 2: MAEs for the nowcasts of year-on-year growth rates of the turnovers indexes, obtained using firm-level data. Results are in percentage points and bolded numbers indicate the lowest MAE for a given estimation period. The bottom row is the MAE of the $t|45$ estimate by the statistical office.

The nowcasting performance of our models does not act homogeneously across the

sectors of interest. For example, factor models are the best, in terms of lower MAEs, for the construction sector, even though the ridge regression performs better at $t|5$. For the services and trade sectors we find that the ridge regression gives the best nowcasts for early estimation periods (i.e., $t|5$ and $t|10$), but when we accumulate more firm-level data, implying a longer publication lag, the best predictions are given by the elastic-net model. It is interesting to notice that the factor model is always doing worse than the shrinkage estimators when we use the final revisions of the turnover indexes, as in Table 1.

Another interesting feature we can infer from Table 2 is how the nowcasting performance of the firm-level data based models reacts to the accumulation of information. We find a fairly large improvement going from $t|5$ to $t|10$, and an even more dramatic drop in MAEs when we go from $t|10$ to $t|20$. Similarly as in Table 1, the new information accumulated 20 days after the end of the reference month does not seem beneficial for our predictions.

Finally, when we compare the performance of our nowcasting models against the degree of accuracy of the first estimates produced by the statistical office (at $t|45$), we obtain a strikingly large improvement in MAE for the services sector already at $t|5$. For the construction sector, the factor models are able to beat the current estimate accuracy at $t|20$. In manufacturing, the ridge regression provides a reasonable performance already at $t|10$, which improves when more data is accumulated, but it is not able to beat the accuracy of the first estimate of the statistical office. In trade, our models perform quite well at $t|20$ (especially the elastic-net method), but are unable to provide more accurate predictions than the statistical office at $t|45$. These considerations highlight the trade-off between timeliness and accuracy for the trade and manufacturing sectors. However, for the services and construction sectors, the faster estimates obtained by using large-dimensional models would even imply an improvement in nowcasting accuracy, which is a remarkable result.

4.1 An extension: improving the accuracy of the first official estimate

The statistical office has major concerns for the quality of their first estimate at $t|45$, which they withhold from publishing to the general public but are produced for internal purposes (e.g., national accounts) and for Eurostat. The turnover indicators growth rates are based on firms' sales, where a share of the observations is obtained from

the turnover inquiry (as described in Section 3, representing a large share of the total turnovers) and VAT data which is obtained from the tax administration (and is available for the total population). Once all the data is collected, Statistics Finland simply sums up all the sales at time t and $t-12$ and compute the year-on-year growth rates, after a careful examination of possible data errors. The base year of the indexes is 2010. A problematic aspect of this approach is that the data sources are available at different time lags. The turnover inquiry is practically complete at $t|28$, but the VAT data is sent to Statistics Finland at $t|52$. Therefore, the first estimate at $t|45$ is based on the observations in the turnover inquiry and imputations, while the $t|75$ (the first official release) incorporates the total population of firms.

We propose a solution to this issue by separating the two data sources and use the information accumulated from the turnover inquiry to forecast (and impute) the year-on-year growth rate of the index of firms which are only available from the VAT. In particular, we use factors extracted from the turnover inquiry to estimate the VAT index growth, assuming that the factors capture economy-wide shocks which affect all Finnish firms, including the ones forming the VAT data. Let's denote the year-on-year growth rate of the turnover index of sector K as y_K. Formally the estimates at $t|28$ are obtained using

$$\hat{y}_{t,K} = \hat{w}_{t,I,K}\left(\frac{I_{t,K}}{I_{t-12,K}} - 1\right) + \hat{w}_{t,VAT,K}\left(\frac{\widehat{VAT_{t,K}}}{VAT_{t-12,K}} - 1\right) \tag{6}$$

In (6), the $\left(\frac{I_{t,K}}{I_{t-12,K}} - 1\right)$ is the change in the index of firms sales in sector K, i.e. $I_{t,K} = \sum_{i=1}^{N} x_{i,t,K}$, that are included the inquiry. $\left(\frac{\widehat{VAT_{t,K}}}{VAT_{t-12,K}} - 1\right)$ is the change in the index of firms of sector K in the VAT subgroup, which is similarly defined as $I_{t,K}$. Finally, $\hat{w}_{t,I,K}$, $\hat{w}_{t,VAT,K}$ are the corresponding weights, computed by $\hat{w}_{t,I,K} = \frac{I_{t,K}}{I_{t,K}+\widehat{VAT}_{t,K}}$ and $\hat{w}_{t,VAT,K} = \frac{\widehat{VAT}_{t,K}}{I_{t,K}+\widehat{VAT}_{t,K}}$.

The unknown elements are the weights of the inquiry and VAT components, and the growth rate of the VAT index (based on firms which are not included in the inquiry or did not respond in time). We use the common factors extracted from the known part of our data (i.e. all the firms included in the sales inquiry, regardless of the sector in which they operate) and use them to forecast the unknown VAT firms index growth. Using the estimated growth and the previous year monthly sum of VAT firms' sales , we compute the current month VAT firms' sales. Subsequently, the current month weights

and the final estimate are obtained. We include the estimated factors in an ARIMA model with automated selection procedure (Proc X12 in SAS) as external predictors, to get fast results. Using this kind of automated approach is an essential requirement for a real-time application, because we have to provide quick results for a large number of series (120 turnover indexes in total). As benchmark, we use an ARIMA model with no factors.

We report the results for the main series we investigated in the previous analysis by using up to two factors as inputs in the ARIMA model. We are able to simulate the actual situation faced by Statistics Finland from January 2015 to September 2016 (because it is the only period for which we can replicate the realistic data accumulation for all 120 sectors), and can compare the results of our estimates of the turnover indexes growth, imputing the unknown VAT component in equation (6), against the current first estimate ($t|45$) of Statistics Finland. The accuracy of an early estimate is computed using the MAE, with the final revision (or the most recent revision) of the turnover indexes growth as target. The results in Table 3 indicate that using the factors extracted from the sales inquiry leads to a much improved accuracy at $t|28$, when compared to the benchmark estimates at $t|45$. This is a remarkable outcome because we provide more accurate estimates (which follow the accuracy recommendation of Eurostat) in a shorter time. Notice that the inclusion of factors is beneficial for all the sectors of interest. Finally, we repeated this analysis for 120 sectors of the Finnish economy and found that our methodology is capable to create more accurate estimates, with a shorter delay, for 118 out of 120 industries considered[2] .

	Manufacturing	Services	Trade	Construction
MAE of estimate 1 factor	0.498	0.465	0.580	1.547
MAE of estimate 2 factors	0.485	0.453	0.554	3.121
MAE of estimate no factors	1.006	0.988	0.981	2.201
MAE of $t|45$ (current method)	1.675	2.161	0.763	2.905

Table 3: MAEs for the nowcasts of year-on-year growth rates of the turnovers indexes at $t|28$, obtained by predicting the VAT component with the accumulated firm level data, using 1 and 2 factors as inputs in an ARIMA model. Results are in percentage points.

[2]For the sake of brevity we do not report the results here but they are available upon request

5 Conclusions

We have examined the nowcasting performance of large dimensional models for the year-on-year growth rate of the turnover indexes, considering the main sectors of the Finnish economy. We find that our firm-level data based specifications are able to beat the ARIMA benchmark, even with a fairly limited amount of data. Moreover, we show that shrinkage methodologies and factor models provide timely estimates of turnover indexes, compared to the first estimates of the statistical office. These methods provide an especially good performance for the service sector, where we beat the official estimates already at the first estimation round (i.e., 5 days after the end of the reference month). Finally, we show that using the factors extracted from a large (albeit incomplete) dataset of firm sales can be helpful in substantially improving the current official estimates of the turnover indexes, by predicting a part of the source data, both in terms of accuracy and timeliness.

References

Filippo Altissimo, Riccardo Cristadoro, Mario Forni, Marco Lippi, and Giovanni Veronese. New eurocoin: Tracking economic growth in real time. *The Review of Economic and Statistics*, 92(4), 2010.

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

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, January 2002.

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

Paolo Fornaro. Predicting Finnish economic activity using firm-level data. *International Journal of Forecasting*, 32(1):10–19, 2016.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1–20, 2010.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: the real time informational content of macroeconomic data releases. *Journal of Monetary Economics*, 55(4), May 2008.

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

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97: 1167–1179, December 2002b.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2): 301–320, 2005. ISSN 1369-7412.