Nowcasting Finnish Real Economic Activity: a Machine Learning Approach

Paolo Fornaro*, Henri Luomaranta**
*Research Institute of the Finnish Economy
**Statistics Finland and University Of Toulouse 1

May 2018

Abstract

We develop a nowcasting framework based on micro-level data in order to provide faster estimates of the Finnish monthly real economic activity indicator, the Trend Indicator of Output (TIO), and of quarterly GDP. In particular, we rely on firm-level turnovers, which are available shortly after the end of the reference month, to form our set of predictors. We rely on combinations of nowcasts obtained from a range of statistical models and machine learning methodologies which are able to handle high-dimensional information sets. The results of our pseudo-real-time analysis indicate that a simple nowcasts' combination based on these models provides faster estimates of the TIO and GDP, without increasing substantially the revision error. Finally, we examine the nowcasting accuracy obtained by relying on traffic data extracted from the Finnish Transport Agency website, and find that using machine learning techniques in combination with this big-data source provides competitive predictions of real economic activity.

1 Introduction

We live in a data-rich world. Statistical agencies, central banks, research institutes and private businesses have access (and produce) thousands of economic and financial indicators. The list of available data is continuously growing, with the introduction of "big data" encompassing sources such as Internet search engines, social media sites, cash registry data and many more. However, this wealth of information has not been directly translated into a faster and more accurate production of important economic statistics, such as the GDP. Statistical institutes publish economic indicators with considerable lag and the initial estimates are revised considerably over time. In Finland, the first estimate of GDP provided by Statistics Finland is released 45 days after the end of the reference quarter (flash estimate), while the first "appropriate" version is released 60

days after the end of the quarter.

The advantages of having a timely picture of the state of the economy are multiple and concern a range of economic actors such as the central bank, the government and private investors and businesses. Providing this type of information in a timely manner would be invaluable, because it would contribute in reducing the uncertainty of the current state of the economy, thus leading to better informed decisions. The economic advantages of having a timely picture of the economy have not been disregarded by the statistical and academic community.

Nowcasting and the production of economic activity indicators in real time have been the focus of a growing literature. Early works related to the tracking of economic conditions in real time are Aruoba, Diebold, and Scotti (2009), for the U.S. economy, and Altissimo, Cristadoro, Forni, Lippi, and Veronese (2010) for the Euro Area. In these studies, the authors develop econometric frameworks with the objective to create high-frequency indicators of real economic activity. On the other hand, the nowcasting literature is interested in estimating an existing economic indicator (usually quarterly GDP growth) in real-time. Few examples drawn from the nowcasting literature are Giannone, Reichlin, and Small (2008), Evans (2005), Modugno (2013), Aastveit and Trovik (2014), among many others. Usually, nowcasting models involve the use of a wide array of data from various sources and different frequencies, such as consumer surveys, financial variables and macroeconomic indicators, and use factor models or large bayesian vector autoregressions to produce predictions of the variables of interest.

In this study, we combine micro-level datasets and machine learning techniques to provide faster estimates of Finnish real economic activity, both at the quarterly and monthly frequencies. In addition, we examine the predictive power of a novel dataset based on traffic volumes' measurements, created by combining disaggregated data obtained from the Finnish Transport Authority website. The use of novel data sources, such as firm-level data and traffic measurements, in combination with the use of a wide array of machine learning techniques provides the main contribution of our study to the nowcasting literature. The use of firm-level data in providing fast estimates of real economic activity is not unique: Matheson, Mitchell, and Silverstone (2010) rely on qualitative responses obtained from business surveys, to obtain nowcasts of New Zealand GDP growth, while Fornaro (2016) uses a similar firm-level dataset to estimate Finnish economic activity. We expand the latter work in two main ways:

firstly, we consider an additional data source, i.e. the trucks' traffic volumes, which can be interesting with respect to the use of big data in economic forecasting and nowcasting (e.g., see Baldacci, Buono, Kapetanios, Krische, Marcellino, Mazzi, and Papailias, 2016). Moreover, we consider a much larger array of statistical frameworks and machine learning techniques, compared to Fornaro (2016), which focuses exclusively on factor models. We show that the machine learning approach is more suitable for modeling this data.

We find that our approach of combining predictions obtained by using a large set of machine learning algorithms, based on firm-level data, is able to provide accurate estimates of monthly economic activity growth, producing revision errors that are in line with the ones of Statistics Finland, while shortening the publication lags by 30 days. The resulting early estimates of the monthly indicator are used to compute nowcasts of GDP year-on-year growth. We provide three early predictions of GDP: the first two are produced during the second and third month of the reference quarter (nowcasts), while the last estimate is computed 16 days after the end of the quarter (backcast). The first two nowcasts provide good accurancy, even though there are some notable revision errors. The estimates produced after the end of the quarter are very accurate, while providing a 45 days reduction in the publication lag. Moreover, the methods we use are computationally feasible and easily automatable, making them appropriate for a real-time setting. We conduct a similar analysis using truck traffic volumes' measurements, and find satisfactory results that, while qualitatively not as good as the ones obtained with firm-level information, allow an even more timely estimation of the economic indicators of interest.

The remainder of this paper is divided as follows: in Section 2 we discuss some of the large set of models adopted in the analysis, in Section 3 we describe our target indicators and data sources. In Section 4, we delineate the structure of our nowcasting exercise, wile we look at the empirical results in Section 5. Finally, Section 6 provides the conclusions.

2 Methodological Aspects

Given the large set of models we employ, an in-depth methodological description is not feasible. However, in this section we try to give the basic intuitions underlying the main classes of models used in this study. The interested readers will be directed to the

original works in which the models we employ were originally developed. Firstly, we look at the (dynamic) factor model. Subsequently, we describe a number of shrinkage methodologies which treat the predictors in a linear manner. Finally, we list some of the more advanced machine learning methodologies that have been working particularly well in our setting.

Before we introduce the specific models, it is important to mention one of the common features that underlies them, i.e. that they are designed to handle large dimensional datasets. A standard statistical model, say the linear regression, cannot handle more than a handful of variables. For example, let's assume that we want to predict the variable y, which includes T observations, using a set of predictors X, of dimensions $T \times K$. In a typical linear regression setting we would fit a model such as:

$$y = X\beta + \epsilon, \tag{1}$$

where ϵ is a normally and independently distributed error term. It can be shown that the variance of the ordinary least squares (OLS) estimate of β, denoted as $\widehat{\beta}$ depends positively on the number of predictors. When K becomes larger the model tends to overfit the in-sample data, which leads to very poor out-of-sample predictions. Moreover, model (1) cannot be estimated using OLS if $K>T$, which is a typical situation we face in our application. Fortunately, the statistical and econometric literatures have developed a series of methodologies that solve the *curse of dimensionality* by using a number of different approaches.

2.1 Factor Models

The main idea underlying factor models is that a small number of constructed variables, factors, can summarize most of the information contained in a large dataset. This approach, together with principal component analysis, has a long tradition in statistics and econometrics. Principal component analysis was introduced by Pearson (1901) and Hotelling (1933), and it has been adopted in a wide range of applications, in psychology, engineering and economics, among others.

Dynamic factor models were introduced in the econometric literature by Sargent and Sims (1977) and Geweke (1977). These first contributions were used in rather small dimensional applications. The introduction of dynamic factor model in large dimensional economic applications is due to Stock and Watson (2002a,b) and Forni,

Hallin, Lippi, and Reichlin (2000). Since these seminal papers, factor models have been adopted in numerous applications and are now an established technique in economic research and policy making.

Let X_t be again $K \times 1$ vector containing our large set of variables a time t. The dynamic factor model specification expresses the observed time series using an unobserved common component (and possibly its lags) and an idiosyncratic component

$$X_t = \lambda(L) f_t + u_t. \tag{2}$$

In model (2), f_t is the $q \times 1$ vector of dynamic factors, u_t is the $K \times 1$ vector of idiosyncratic components, L is the usual lag (backshift) operator and $\lambda()$ is the $K \times q$ matrix of factor loadings. The dynamic factors are modeled following

$$f_t = \Psi(L) f_{t-1} + \eta_t, \tag{3}$$

where $\Psi(L)$ is $q \times q$ lag polynomial. The idiosyncratic disturbances in (2) are assumed normal and uncorrelated with the factors at all leads and lags. In the exact factor model, u_t are assumed to have no autocorrelation or cross-sectional correlation (i.e. $E(u_{it}, u_{jt}) = 0$ for $i \neq j$), while the approximate factor model allows for mild auto and cross-sectional correlation.

If the lag polynomial $\lambda(L)$ has finite order p, then (1) can be rewritten

$$X_t = \Lambda F_t + u_t, \tag{4}$$

where $F_t = [f_t', f_{t-1}', \ldots, f_{t-p+1}']$ is $r \times 1$ and Λ is the $K \times r$ matrix of factor loadings. Representation (4) is the static factor model version of model (2)-(3), in which the r static factor consists of the current and lagged values of the q dynamic factors.

One of the most popular techniques to estimate F_t in (4) is principal components. This estimator is derived from the least squares problems,

$$\min_{F_1,\ldots,F_T,\Lambda} V_r(\Lambda, F) = \frac{1}{KT} \sum_{t=1}^{T} (X_t - \Lambda F_t)'(X_t - \Lambda F_t), \tag{5}$$

subject to $K^{-1}\Lambda'\Lambda = I_r$. The solution to this maximization problem is to set $\widehat{\Lambda}$ to the scaled eigenvectors corresponding to the r largest eigenvalues of $\widehat{\Sigma}_{XX} = T^{-1} \sum_{t=1}^{T} X_t X_t'$. It follows that the least squares estimator of F_t is $\widehat{F}_t = N^{-1}\widehat{\Lambda} X_t$, which are the first r

principal components of X_t. Stock and Watson (2002a) have shown that the principal component estimator of the factors is consistent also in the presence of mild serial- and cross-correlation in u_t.

Static principal components, described in the previous paragraph, have been one of the most used methods to estimate factor models. However, there have been multiple methodologies that have been proposed in the literature. Among them, notable examples are the dynamic principal component of Forni et al. (2000), and the hybrid principal components and state space estimation of Doz, Giannone, and Reichlin (2011). Bai and Ng (2002) developed a series of information criteria that provide an estimate of the number of static factors r which they show to be consistent, assuming the the number of factors is finite and does not increase with (K, T).

2.2 Shrinkage Models

While the factor model described in the previous subsection solves the *curse of dimensionality* by extracting a relatively small number of variables from our large dimensional dataset, resulting in a two-step procedure, shrinkage methodologies regularize the coefficients of the original predictors. Next, we examine three regularized regression approaches, namely the ridge regression, the lasso and the elastic-net. One similarity among these models is that the predictors are included linearly. Later on, we are going to describe approaches that augment the set of predictors with a number of nonlinear transformations.

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

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{K} \beta_j^2, \tag{6}$$

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

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{K} |\beta_j|. \tag{7}$$

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

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda_1 \sum_{j=1}^{K} |\beta_j| + \lambda_2 \sum_{j=1}^{K} \beta_j^2 \tag{8}$$

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 it 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.

2.3 Machine learning approaches

So far, we have described methodologies that, despite being able to solve the curse of dimensionality, assume a linear relationship between the predictors and the target variables. In our study, we have examined the nowcasting ability of a large number of machine learning methods, going from tree-based models to boosting and neural networks. We are not going to offer a thorough examination of these techniques, however we go over the main intuitions and principles underlying the main families of machine learning methods that we have adopted. A much more detailed discussion of these models can be found in Hastie, Tibshirani, and Friedman (2009).

Boosting

Boosting is a form of forward stage-wise modeling, where our target variable of interest y_t can be expressed as an additive function

$$\hat{f}_M(X_t) = \sum_{m=1}^{M} b(X_t, \hat{\beta}_m), \tag{9}$$

for $t = 1, \ldots, T$, where T is the number of observations we have. In (9), $b(X_t, \hat{\beta}_m)$ are called *learner* and are a, possibly non-linear, function of the predictors. M represents the total number of boosting iterations which governs how the final model fits the data. Notice that the boosting procedure is feasible in a high-dimensional setting because for each iteration m the parameters estimated in the previous iteration are left unchanged. Define $\bar{y}$ as the sample average of the target variable and $L(y_t, \hat{f}_m(X_t))$ as our loss function. The general boosting algorithm can be summarized as

1. Set $f_0(X_t) = \bar{y}$.

2. For $m = 1, \ldots, M$

(a) Compute

$$\hat{\beta}_m = \underset{\hat{\beta}}{\operatorname{argmin}} \sum_{t=1}^{T} L(y_t, \hat{f}_{m-1}(X_t) + b(X_t, \hat{\beta}))$$

(b) Set

$$\hat{f}_m(X_t) = \hat{f}_{m-1}(X_t) + b(X_t, \hat{\beta}_m)$$

To give some additional insights on the boosting procedure, assume that our loss functions is the typical squared error loss

$$L(y_t, \hat{f}_m(X_t)) = 1/2(y_t - \hat{f}_m(X_t))^2$$

and that our learner is linear, i.e. $b(X_t, \hat{\beta}_m) = X_t\hat{\beta}_m$. The resulting algorithm can be described:

1. Set $f_0(X_t) = \bar{y}$.

2. For $m = 1, \ldots, M$:

   (a) Compute $u_t = y_t - \hat{f}_{m-1}(X_t)$.

   (b) For $k = 1, \ldots, K$ regress u_t on $X_{k,t}$ to obtain $\hat{\beta}_k$ and compute $SSR_k = \sum_{t=1}^{T}(u_t - X_{k,t}\hat{\beta}_k)^2$.

   (c) Choose $X_{k^*,t}$ which yields the minimum SSR_k.

   (d) Update $\hat{f}_m(X_t) = \hat{f}_{m-1}(X_t) + \nu X_{k^*,t}\hat{\beta}_{k^*}$.

In step (d), ν is a regularization parameter that lies between zero and one. Notice that the algorithm described above will lead to select one additional variable for each step m. One common approach to estimate the total number of boosting iterations M is cross validation, i.e. we divide the original dataset into a number of equal parts. We keep all but one part to estimate the model for a given M and the remaining data are used to evaluate the performance. This procedure is repeated for all splits and the resulting errors are averaged.

While boosting was initially developed as a classification technique, there have been a number of econometric forecasting applications which rely on this model. Two examples are Bai and Ng (2009) and Wohlrabe and Buchen (2014).

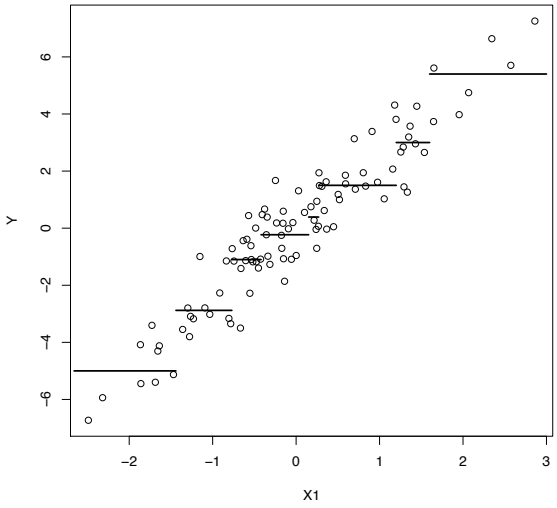
Tree-based methods

Tree-based techniques partition the space of explanatory variables in order to fit a simple model for each partition. To make the idea more clear, let's proceed with a very simple example. Assume that we have a variable Y which is a simple linear function of an individual predictor X plus a normally distributed error. For this kind of scenario, linear regression models would work just fine but this kind of trivial application can be useful to grasp the intuition behind regression trees.

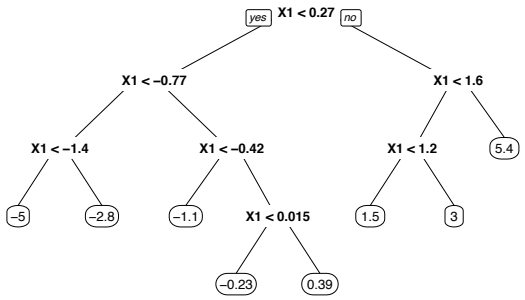
In a basic regression tree, we split the X space in different regions and fit a constant model for each region. Formally, assume that we have P partition of the X space, then we have

$$\hat{f}(X) = \sum_{p=1}^{P} c_p I\{X \in R_p\}, \tag{10}$$

where $I(X, R_p)$ is an indicator function which is equal to 1 if the X belongs to partition R_p. It can be shown that under squared loss function the optimal estimate of c_p is simply the average of Y conditional on X belonging to R_p. In our example, we simulate 100 observations of the aforementioned process and fit a simple regression tree model. The resulting scatterplot and the graphical representation of the tree are reported below.



(a) Scatterplot of the tree-based regression



(b) Graphical representation of the tree-based regression

Figure 1: Regression tree example

Figure 1 (b) gives a fairly clear representation of the regression tree technique. It separates our X values into 7 regions (corresponding to seven splits). For each region,

we calculate the average Y so that when we get a new value of X the corresponding predictions $\hat{Y}$ will simply be the average of Y corresponding to region p to which the new X belongs. For example, if our new X is between -0.42 and 0.015, then the predicted Y will be -0.23.

Naturally, regression trees can deal with numerous predictors, which will then impact the optimal splits and tree size. Notice that we need to estimate the optimal splits of the predictors and the depth of the tree. While the first aspect is estimated, how large we should grow the tree is left as a tuning parameter. A typical strategy is to grow a very large tree and then prune the tree afterward, to reduce its complexity.

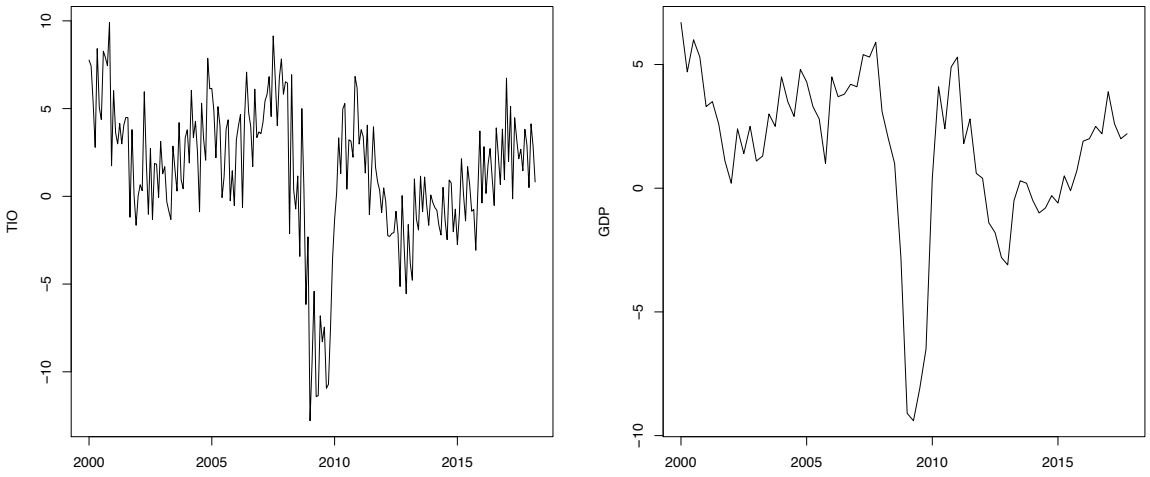
3 Data description

The main predictors in our nowcasting application are firm-level sales extracted from the sales inquiry, a monthly survey conducted by Statistics Finland for the purposes of obtaining turnovers from the most important firms in the economy. This dataset covers around 2,000 enterprises and encompasses different industries (services, trade, construction, manufacturing), 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 15 to 20 days after the end of the reference month. Formally, Statistics Finland imposes a deadline to the firms, which are supposed to send their data by the end of the 15th day of the month. We compute the nowcast on the 16th day. However, this deadline is not always met, thus our set of firms' sales does not cover the entire sample. The data accumulation is realistically simulated by using the time stamp of the reported sales, which allows us to track what data was available by each date of a month. Further, the more recent data points, starting from January 2017, are based on real time data collection.

A similar set of explanatory variables 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 2006), and that they have reported sales figures by the date we extract their information from the database. We collect data of the firms that have reported the sales by 16 days after the end of the reference month because it is right after the deadline for enterprises to send their figures. This choice leads us to have 800 firms on average, in the predictors' set. We compute the sales growth rates for all the

months from 2006 until the nowcasted month of interest. If the firm has reported sales by the $t + 16$ at the nowcasted month, but has missing values during the time span (i.e. the firm did not reply at some earlier date, or the firm was not included in the turnover inquiry at that time), we try to obtain the missing growth rates from VAT data, which should include all the firms in the economy. Notice that our resulting data does not contain missing values.

The target variables in our exercise are the Trend Indicator of Output (TIO) and quarterly GDP, both measured in real-term year-on-year growth rates. The TIO is a monthly series that describes the development of the volume of produced output in the economy. It is constructed by using early estimates of turnover indexes (not publicly available), which are appropriately weighted to form the monthly aggregate index. The TIO is published monthly at $t + 45$, and its value for the third month of a quarter is used to compute the flash estimate of GDP, which is also published as an early version at $t + 45$, and updated at $t + 60$. The $t + 60$ version is considered as the first official and reliable estimate of GDP. Thus, given the information we have provided, the TIO in fact represents a GDP nowcast in its own right. We stress the importance of using the realistic vintages, as the data is typically "improved" by many internal processes, and by the accumulation of new data. The usage of revised data can arguably lead to too optimistic views on the nowcasting performance. We have been very careful about this point, and are therefore convinced that the test results we present provide an accurate estimate of the accuracy of a real-time application. Below we report the plots of the TIO and GDP year-on-year growth rates.



(a) TIO year-on-year growth, monthly series

(b) GDP year-on-year growth, quarterly series

Figure 2: Target variables

One aspects that it is important to underline is how closely related the TIO and GDP growth are. If we aggregate TIO growth to the quarterly level we obtain a series that closely tracks GDP growth (the resulting correlation coefficient is 0.99). This demonstrates that providing a good estimate of TIO leads to a greater nowcasting accuracy of GDP.

3.1 Traffic data

Big data sources provide interesting possibilities for nowcasting, given that they are collected real-time, in an automated manner. The firm-level data which constitutes the main data source of our exercise provides a good nowcasting performance, as we are going to show in Section 4. However, while high-dimensional, the firm-level turnovers we use are not a big data source in a traditional sense, even though they have some similar characteristics, namely that they represent an incomplete and not necessarily a representative set of information which gradually accumulates as time passes. The key difference, in the Finnish setting at least, lies in the real-time availability of the data, since the firms start sending information only after the reference month has ended. Moreover, our turnover dataset is structured and fairly easy to handle, which is not typical of big data.

We examine traffic loop data for real-time estimation purposes, and consider the predictive performance of traffic volumes records obtained by the Finnish Transport

Agency website[1]. This dataset contains the number of vehicles passing through a number of measurement points (about 500) around Finland, observed through an automatic traffic monitoring system. The data is available at hourly frequency, and it distinguishes between different types of vehicles. This dataset contains numerous missing values, due to the fact that some measurement points do not have observation for certain days or months, and it is not structured. For our nowcasting analysis, we collect data for trucks' traffic volumes from January 2010 (the first dataset available), in particular their year-on-year growth rate at the different measurement points across the country. Trucks' traffic presents an interesting intuitive link with aggregate economic activity. We expect that in periods of economic growth, when trade volumes, and production are increasing, we should observe a higher number of trucks' passages, in order to move goods. Of course, this does not cover the transfer of services and other types of economic activities, but it should still present some positive correlation with economic activity growth. More details around how we implement this data source in our predictive framework is provided in Section 4.2.

4 Nowcasting Finnish economic activity

A nowcasting technique is of little use if it cannot be applied in a real setting, which has been the key motivation for conducting this study. This is why we have been extremely careful in setting up our testing procedures, and collecting the original vintages of the data sets, as we will explain in this section.

4.1 Nowcasting exercise formulation

To make sure that the overall nowcasting procedure is feasible in a real-time setting, we need to consider two important aspects: data availability and computational feasibility. The first issue boils down to the fact that, while testing the nowcasting models, the researcher should not rely on data which would not be available in real time. This implies that we have to take into account the publication lag in a realistic fashion. For example, in our application we compute the nowcast at $t + 16$, i.e. 16 days after the end of the reference month or quarter, thus we should not use data sources which are not available by then (for example VAT data). The other important aspect revolving around data availability concerns the use of the correct vintage of data, which is the

[1]The data is available at https://aineistot.liikennevirasto.fi/lam/reports/LAM/ .

one that reflects the information available at the time the nowcast would have been computed. Most economic series are revised multiple times, both because of estimation error and of benchmarking. The practitioner should avoid using the final value of the indicators of interest, including the target variables (in our case, GDP and the TIO) and the predictors, and focus on collecting realistic, non-revised, versions of the indicators.

In our nowcasting exercise, we are careful in terms of making a realistic representation of the available information set. With respect to the target variables, things are rather straightforward. Computing estimates at $t + 16$ means that we have the previous value of TIO. For example, suppose that we want to nowcast TIO for March: we would compute the nowcast on April 16th and, given the release schedule, at that date we have TIO data for February. We would then estimate a model using data up to February and then compute the nowcast using the March firm-level sales. When estimating quarterly GDP, we do not rely directly on the GDP series but rather use TIO, which means that we do not have problems in terms of publication lag. Fortunately, we are able to realistically simulate the accumulation of firm-level data, because Statistics Finland records that date on which the firms send their sales reports.

While the publication lags are easy to take into account in our setting, the use of realistic vintages has proven to be somewhat harder to tackle. The TIO is revised multiple times, even many months after its release. Moreover, these revisions do not incorporate solely corrections of the estimate due to the expansion of the data sources but are also affected by benchmarking. This fact implies that if we use the final version of TIO in the estimation and in the evaluation of the nowcasts we would put ourselves in a dramatically different scenario than the one faced in real time by the statistical office, who we assume is interested in producing the nowcast. Moreover, our nowcasts would contain errors that are not due to the lack of predictors but that are instead caused by the lack of smoothing and benchmarking. Consequently, we use vintages reflecting the first estimate of TIO and adopt these initial figures as target to evaluate our nowcasts. Unfortunately, the historical vintages for TIO are available only since March 2012, meaning that our nowcasting exercise does not cover some interesting periods such as the Great Recession of 2008–2009. However, we are left with more than 60 predictions to be made and the timespan going from 2012 until the beginning of 2018 does include periods of high growth and months of considerable output drop. On the predictors' side of things we a have a similar problem, i.e. the firm-sales are revised

over time. These corrections include actual revisions made by the firms (even though these adjustments are relatively small) and the corrections for organic growth made by Statistics Finland. In particular, the statistical institute adopts a growth-correction methodology which cleans sales growth caused by mergers and acquisition. While the timing of these corrections is not clear, we want to avoid being overly optimistic in terms of the data availability at the time of the nowcast, thus we rely on the original, not corrected, version of the firm-level data.

Now to the structure of our empirical exercise: we start to compute monthly nowcasts of the TIO from March 2012. In particular, we extract a panel of firm-level sales which starts from January 2006 and contains information until March 2012. Notice that our panel is balanced (i.e. we select firms which are present throughout the time interval of interest). In real-time setting, this nowcast would have been computed in April 2012, specifically 16 days after the end of the month we nowcast. The models are estimated using the vintage of TIO available in April 2012. We repeat this procedure for each month until March 2018, expanding the estimation window (instead of using a rolling window approach). This means that our estimation sample is increasing over time. As an example, in the case where we use the estimated factors as predictors we would summarize our procedure as:

$$y = \hat{F}\beta + \epsilon \tag{11}$$

$$\hat{y}_t = \hat{F}_t\hat{\beta} \tag{12}$$

In (12) and (13), t refers to the month we want to nowcast and y and $\hat{F}$ are the TIO and estimated factors going from $t = 1, \ldots, t-1$. Of course (12) and (13) take many forms depending on the model we adopt, but the principle is similar: we first estimate the models using data until the latest month for which we have TIO values and then we use the most recent firm-level information to compute the nowcast, given the estimated model parameters.

Our quarterly estimate of GDP are entirely based on TIO, both the released version and our nowcasts. As we mentioned in the data description, TIO provides the basis for the initial estimate of GDP, hence it is optimal to use it as a predictor in a nowcasting exercise. We compute the GDP nowcasts differently, depending on the month in which we make the estimate. In our setting, the nowcasts for a given quarter are computed three times: during the second month of the quarter, during the third month and 16

days after the end of the quarter. In the first case, we would use the nowcast of TIO for the first month of the quarter, then estimate an automated ARIMA model (see Hyndman and Khandakar, 2008) to obtain the forecasts of the remaining months. If we compute the GDP nowcast during the third month, we would use the first TIO estimate made by Statistics Finland for the first month, then use our nowcast of TIO growth for the second month and then compute the 1-step ahead forecast for the third month. When we estimate GDP growth 16 days after the end of the quarter we use the TIO growth computed by Statistics Finland for the first two months and augment them with our nowcast of TIO for the last month of the quarter. Eventually, we are going to have an estimate of TIO growth for each month of the quarter of interest and we obtain GDP growth by taking a simple average over the three months. Denote the estimate of GDP growth for quarter q going from to month $t-2$ to t as $\widehat{GDP}_{q,t}$, then our quarterly nowcast is $\widehat{GDP}_{q,t} = 1/3(\hat{y}_{t-2} + \hat{y}_{t-1} + \hat{y}_t)$ Notice that this procedure is rather similar to the one of bridge regression, which links quarterly and monthly variables via simple linear models. We have tried to estimate a linear regression of GDP growth onto the quarterly average of TIO growth, i.e. estimating the linear model $\widehat{GDP}_{q,t} = \beta \frac{(\hat{y}_{t-2}+\hat{y}_{t-1}+\hat{y}_t)}{3} + \epsilon_t$, but our results indicate that the simple average of TIO growth is a better predictor than using the bridge formulation.

The other issue that we mentioned at the beginning of this subsection concerns computational feasibility. We estimate more than 150 nowcasting models, some of which are computationally burdensome. Given that we would like to produce (and possibly release) the nowcasts around $t + 16$, using the information set available by then, we need to find some sort of compromise between having the largest spectrum of models and being able to estimate TIO quickly. In order to do that, we select a relatively small subset of models (around 20) which perform well on the historical sample and proceed to use these techniques to produce nowcasts for the most recent month. We then average these nowcasts using simple combination schemes such as unweighted average or using weights which depend on historical nowcasting performance (Stock and Watson, 2004, point out that these schemes outperform more complex ones). We have tried different criteria in order to trim the original nowcasting models and found that keeping the models with lowest mean error (i.e. the ones producing unbiased nowcasts of TIO) tend to produce the best TIO and GDP estimates, once combined. One we have produced the fast estimate of the indicator of interest, we re-evaluate the whole

set of models to make sure that the performance with respect to the latest months does not alter the best set of models. This implies that, in principle, the models which are going to be included in the estimate can change over time.

4.2 Nowcasting with traffic measurement data

As we mentioned in Section 3.1, traffic volumes data represent a more complicated data source compared to our set of firm-level sales. For example, they present many missing observations and the panel of measurement points needs to be constructed from the original files available on the Finnish traffic authority's webpage. Given that the data is available only from January 2010, we have decided to start the computation of pseudo-real-time nowcasts of TIO growth from January 2014, to give us four years of estimation sample. Similarly as in the firm-level data case, we adopt the predicted TIO growth rates to compute the year-on-year growth of GDP.

The traffic data is aggregated at the monthly level and we assume that our estimation of TIO is conducted around 16 days after the end of the reference month (as in the main exercise). This allows us to use the Statistics Finland's estimates of TIO for the $t-1$ month, where t represents the period we want to nowcast. However, in principle the traffic data we utilize allows for nowcasts during the month of interest, given their daily frequency. It is important to point out that, unlike the firm-level data we utilize, our set of traffic volumes contains missing values. In order to impute the missing observations, we rely on the regularized principal component technique illustrated in Josse and Husson (2016).

The actual nowcasts are computed using statistical models and machine learning techniques similar to the ones described in Section 2. The final nowcasts are obtained by making a simple unweighted average of the individual predictions, after trimming the modes producing large historical mean errors.

5 Empirical results

5.1 Results for TIO nowcasts

As pointed out in Section 3, the TIO is a monthly indicator of real economic activity. Our nowcasting exercise is centered on providing fast estimates for the year-on-year growth rate of TIO, starting from March 2012 (the first month for which we have the

vintage of the data) and ending in March 2018. We now provide the results for our pseudo out-of-sample analysis. Specifically, we report the results of the models which provide the lowest root mean squared error (RMSE), the lowest mean error (ME), mean absolute error (MAE), and finally for the model with the lowest maximum absolute error (MaxE). In addition, we report the results for the simple forecast combination consisting of the unweighted average of the nowcasts provided by the 20 models with lowest MEs[2]. This choice is driven by the high importance, for the statistical institute, of having unbiased flash estimates. We plot the nowcasts obtained from the forecast combination, against the first published version of TIO.

[2]This set includes specifications from the regressions trees class, random forests, factor models, ridge regression, regression splines and k-nearest neighbors.

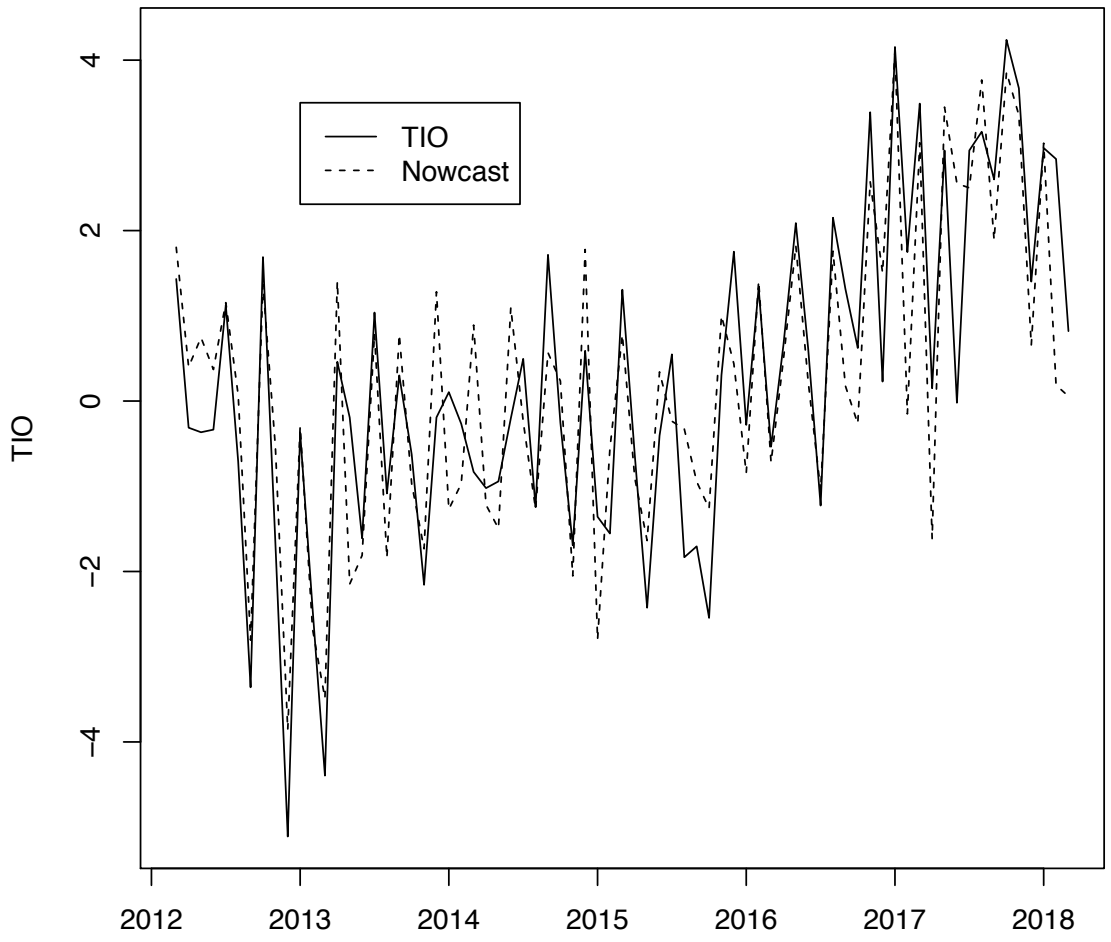


Figure 3: First version of TIO year-on-year growth and nowcasts combination, using the unweighted average of models selected based on low mean errors. The first version of TIO is published 45 days after the end of the reference month, while the nowcasts are computed 16 days after the end of the reference month. The set of predictors is based on firm-level turnovers.

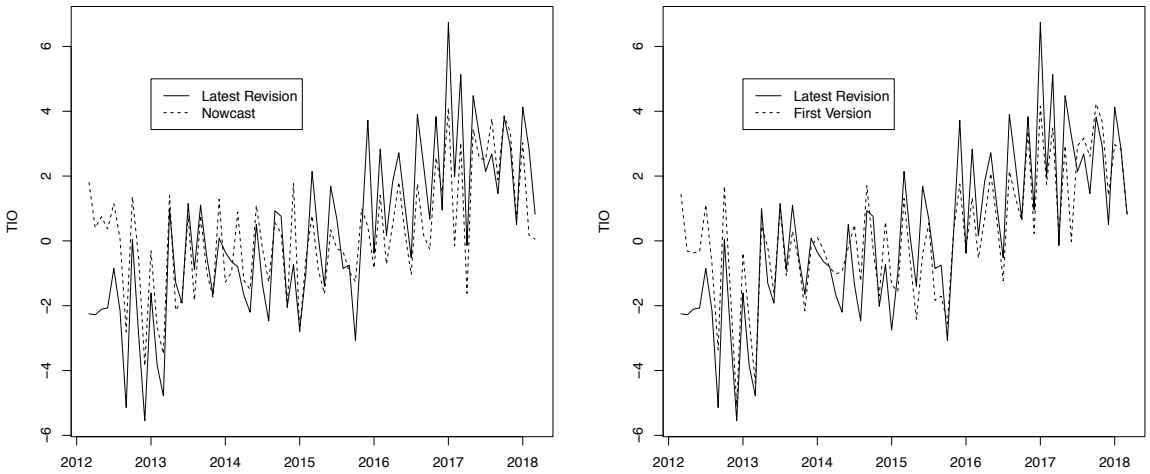
Plots are not the most accurate tools to evaluate the performance of a nowcast model, but they do provide some intuition on the usefulness of our predictions. In this case, it seems that our firm-level data provides a good basis for providing flash estimates of TIO. The nowcasts track fairly well the original series, except for a fairly large mistake in April 2017, while they provide a substantial gain in terms of publication lag (around 30 days). Next, we provide some numerical indicators of the nowcasting performance, for the models described at the beginning of this subsection. Moreover, we report the results obtained by using an automated ARIMA procedure, using the latest available TIO vintage at the time of the nowcast.

	Lowest ME	Lowest RMSE	Lowest MAE	Lowest MaxE	Combination	ARIMA
ME	0.00	-0.06	0.04	0.03	-0.00	0.23
MAE	1.05	0.82	0.81	0.82	0.76	1.15
RMSE	1.29	1.03	1.1	1.05	0.95	1.46
MaxE	3.8	2.8	4.3	2.5	2.65	3.6

Table 1: ME, MAE, RMSE and MaxE for different nowcasting models. Lowest ME, RMSE, MAE and MaxE indicate the models with the lowest mean error, root mean squared error, mean absolute error and max error, respectively. The Combination column contains performance measures for the simple nowcast combination based on the unweighted average of our models. The set of predictors is based on firm-level turnovers.

As we can see from Table 1, the nowcasting performance of our selected models is better than the one of an automated ARIMA procedure. Moreover, the simple nowcast combination provides the best estimates, in terms of ME, RMSE and MAE. However, the largest error of the combination is slightly larger than the one of the lowest MaxE model. In our case, nowcast combinations seem to be the most desirable approach, also in the light of being less prone to possible structural breaks in a model's performance. Consequently, for the rest of this paper, e.g. when we look at the results for quarterly GDP growth, we focus on the nowcasts obtained by combining different model predictions.

The main target of our nowcasts is the first version of the TIO. This is because the later versions of this series are adjusted both for prediction errors and for additional benchmarking, meaning that we cannot be sure whether the nowcast error is due to the mistake in the prediction or because of some subsequent benchmark. However, it is still interesting to check the performance of our nowcasting framework against the final version of TIO, also because it allows us to compare our revision error against the one based on Statistics' Finland publications. We first plot the nowcasts obtained by combining the original predictions, together with the latest version of TIO. We also plot the first version of TIO against the final revision available.



(a) TIO year-on-year growth, final version and nowcasts combination.

(b) TIO year-on-year growth, final version and first publication.

Figure 4: TIO year-on-year growth rate, first publication, final version available and nowcast. The set of predictors is based on firm-level turnovers.

Figure 5 (a) shows a lower nowcasting performance for our approach, which is expected, given that the TIO series we use in the estimation of our model has substantial difference from its later revisions. This can be seen from Figure 5 (b), where we depict the first and final version of TIO: the difference between the two series is remarkable, especially for certain periods. For example, the first official release of the year-on-year growth of TIO for June 2017 was -0.02 percentage point, which was then revised to 3.25 percentage points (interestingly, our nowcast for this month is much closer to the final value of TIO than the first release of Statistics Finland). While such extreme revisions are not common, they do show the difficulties in creating flash estimates of real economic activity. Next, in Table 2, we report the predictive performance measures for the nowcast combination approach, using the final value of TIO as target, even though we still use the original vintages of TIO in the estimation. We also report the same measures to evaluate the performance of the Statistics Finland's first publication.

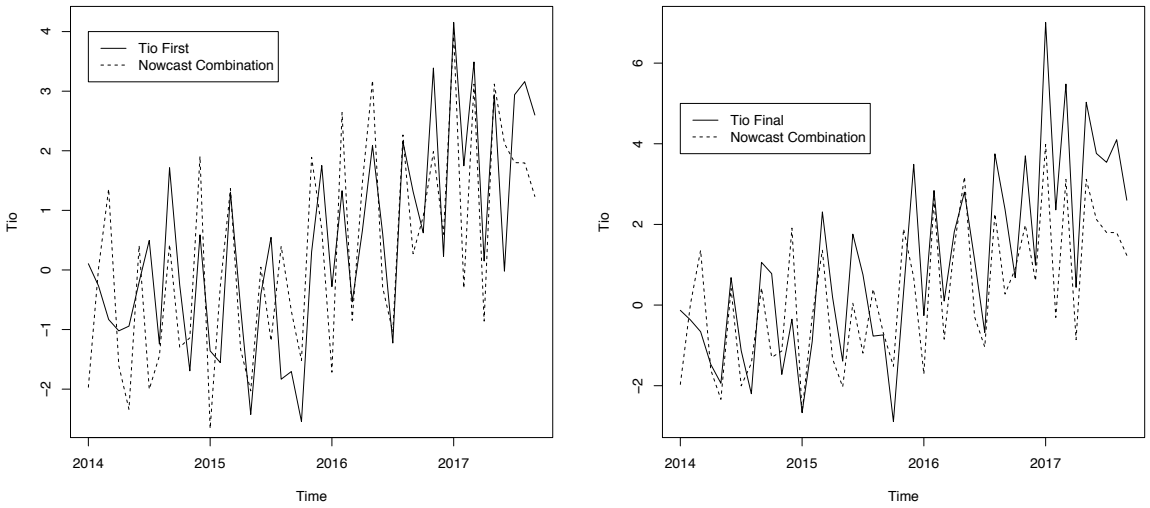
	Combination	Statistics Finland's first
ME	-0.01	-0.004
MAE	1.12	0.92
RMSE	1.38	1.14
MaxE	3.26	3.27

Table 2: ME, MAE, RMSE and MaxE for the nowcast combination approach and for the Statistics Finland's first publication of TIO. The target is the latest available version of the year-on-year growth of TIO. The set of predictors is based on firm-level turnovers.

The performance measures reported in Table 2 confirm the fact that our nowcasting approach fairs worse when it is evaluated using the latest revision of TIO. However, it is interesting to see that the predictions of our simple nowcasting combination do not show a much larger revision error compared to the first publication of Statistics Finland (which suffers from a much longer publication lag), especially when considering the maximum absolute error.

So far, we have evaluated the performance of nowcasts based on firm-level turnovers, the core predictors of this study. However, as mentioned before we have also constructed flash estimates based on measurements of trucks' traffic volumes[3]. First, we report the plots of the predictions obtained by simple model combinations, where we exclude the models with historically large mean errors. We depict both the nowcasts against the first version of TIO and compared to the latest available revision.

[3]The results concerning months after September 2017 were computed very recently, using firm-level data. Given that the traffic data is of secondary interest for this study, we did not compute nowcasts based on that data source after September 2017.



(a) TIO year-on-year growth, first version and nowcasts combination.

(b) TIO year-on-year growth, final version and nowcast combination.

Figure 5: TIO year-on-year growth rate, first publication, final version available and nowcasts. The set of predictors is based on trucks' traffic volumes.

While there are still some substantial nowcasting errors, it is impressive that an unstructured and peculiar data source such as traffic volumes is able to provide estimates that track economic activity fairly well. To gain a better grasp of how our approach is performing, we report the nowcast error measurements that we have used throughout the report, both for the first and final version of TIO.

	Combination vs. First	Combination vs. Final
ME	-0.15	-0.73
MAE	1.02	1.24
RMSE	1.21	1.48
MaxE	2.50	3.02

Table 3: ME, MAE, RMSE and MaxE for the nowcast combination approach, evaluated using the first version of TIO growth and its latest available version.The set of predictors is based on trucks' traffic volumes.

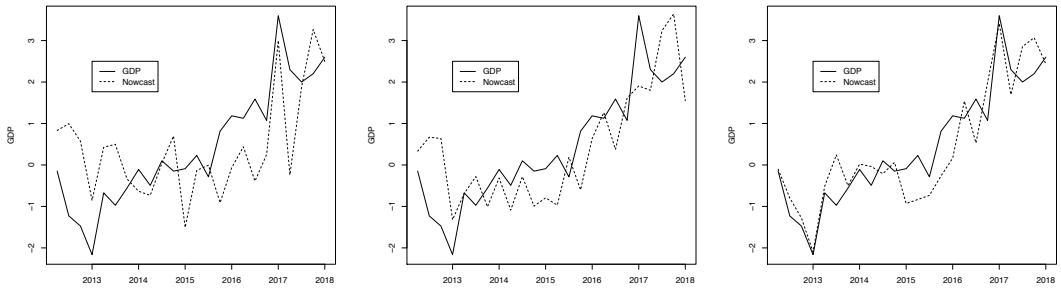
Table 3 gives us some really interesting insights. With respect to the first version of TIO, the nowcasts combination based on traffic data provides slightly worse predictions, at least compared to the sales' data. However, the MAE and MaxE are fairly low, indicating a satisfactory nowcasting performance. When looking at the results for the latest revision we find a surprisingly small maximum absolute error, even smaller then the one of Statistics Finland's publication. Moreover, the mean absolute error and

mean squared error are very similar to the ones of nowcasts based on firm-level data. The main issue with traffic data is the presence of a fairly large (in absolute terms) mean error, indicating that our nowcasts are biased with respect to the latest version of TIO. However, we have to keep in mind the nowcasting errors obtained from the comparison with the latest revision of TIO might be caused, partially, by smoothing or benchmarking that cannot be predicted.

To summarize the results of this subsection, we have seen that combining firm-level data with statistical models and machine learning techniques that are able to deal with large dimensional datasets provide fairly accurate nowcasts, both with respect to the first and to the final version of TIO. The good predictive performance is matched with a substantial gain in timeliness, around 30 days compared to the current publication schedule. The results for the estimates based on traffic volumes evidence the potential of this data source. While the predictions are slightly worse then the ones based on firm-level data, especially compared to the first release of TIO, the errors are not extremely large. Notably, the maximum revision error obtained from this data source is even lower than the one of the first Statistics Finland's publication. The potential real-time availability of traffic data, combined with their satisfactory nowcasting performance, indicates that it is a data source that should be studied further.

5.2 Results for quarterly GDP nowcasts

We now turn to the results regarding the estimation of quarterly GDP year-on-year growth, in real terms. In particular, we nowcast the $t + 60$ release of GDP, which is the first official release made by Statistics Finland. In Section 4.1, we describe how we use the nowcasts of TIO to compute GDP growth, while this subsection is devoted to the reporting of the results. As we did for TIO, we start by plotting our nowcasts (again obtained by the simple unweighted average of the original predictions), against the official GDP growth. We do this for the nowcasts computed during the second month of the quarter, the ones produced during the third month and finally the nowcast computed 16 days after the reference quarter. The nowcasts are provided for the period going from 2012 Q2 until 2018 Q1 (the last observation of GDP is actually based on the flash estimate provided by Statistics Finland, instead of the $t + 60$ release).



(a) Nowcasts second month (b) Nowcasts third month (c) Nowcasts 16 days after

Figure 6: GDP year-on-year growth rate, first publication and nowcasts obtained with simple unweighted average of the predictions. The set of predictors is based on firm-level sales.

Figure 6 indicates that the estimates of TIO based on our nowcasting approach provide good predictions for GDP growth, in a timely fashion. The performance of our models seem to be particularly strong when we compute the predictions during the third month of the quarter and 16 days after the end of the quarter, providing us a 45 to 75 days reduction in the publication lag. Next, we report the nowcasting performance measures for these three sets of predictions. We also compare our results against the forecasts obtained by using an automated ARIMA process for quarterly GDP.

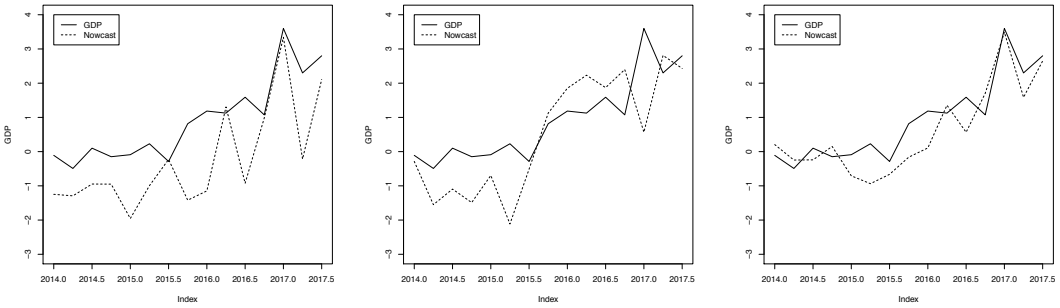
	Nowcast second month	Nowcast third month	Nowcasts 16 days after	ARIMA
ME	-0.04	-0.03	-0.03	0.16
MAE	0.99	0.86	0.53	0.95
RMSE	1.22	1.02	0.65	1.21
MaxE	2.5	2.1	1.21	2.4

Table 4: ME, MAE, RMSE and MaxE for the nowcast combination approach, evaluated using the first version of quarterly GDP year-on-year growth. The set of predictors is based on firms' sales. Nowcast second month refers to the estimates of GDP computed during the second month of the reference quarter, nowcast third months are the estimates computed during the third month of the quarter and nowcasts 16 days after are computed after the end of the reference quarter.

Looking at Table 4, we see that our nowcasting framework is able to predict GDP accurately. As we can expect, the performance of the models improves the later we compute the nowcasts and, from the second estimate onward, they are able to beat a simple ARIMA benchmark. In particular, the latest estimates, which allow for a 45 days reduction in publication lag, present a very low MAE and a low maximum error. Overall, we can say that the nowcasts of TIO based on firm-level data are a good basis to estimate real economic activity.

Finally, we examine the performance of the nowcasts based on traffic data. We

start by depicting plots similar to the ones in Figure 6, i.e. we report the predictions computed during the second and third month of the reference quarter, together with the 16 days after the end of the quarter estimates. Notice that these nowcasts go from 2014 Q1 until 2017 Q3.



(a) Nowcasts second month. (b) Nowcasts third month. (c) Nowcasts 16 days after.

Figure 7: GDP year-on-year growth rate, first publication and nowcasts obtained with simple unweighted average of the predictions. The set of predictors is based on truck's traffic volumes.

The quarterly results confirm the promising performance of traffic data for the production of early estimates of GDP. However, from the graphs it seems that the estimates computed during the second and third month of the quarter are less reliable than the ones based on firm-level data. On the other hand, the $t+16$ nowcasts track well GDP growth, or at least do not show a substantially different performance compared to the ones obtained through the firm-level sales. To asses in a more formal way the performance of our nowcasts, we report the error measures as before.

	Nowcast second month	Nowcast third month	Nowcasts 16 days after
ME	-1.15	-0.41	-0.32
MAE	1.18	0.97	0.55
RMSE	1.46	1.25	0.65
MaxE	2.52	3.03	1.16

Table 5: ME, MAE, RMSE and MaxE for the nowcast combination approach, evaluated using the first version of quarterly GDP year-on-year growth. The set of predictors is based on trucks' traffic volumes. Nowcast second month refers to the estimates of GDP computed during the second month of the reference quarter, nowcast third months are the estimates computed during the third month of the quarter and nowcasts 16 days after are computed after the end of the reference quarter.

The results of Table 5 confirm the intuition we gathered from Figure 7, i.e. that the nowcasts produced using traffic date have a lower predictive performance compared to the ones based on firm-level sales. This is especially true for the estimates during the second and third months of the quarter. On the other hand, the performance of

the $t + 16$ estimates have a similar nowcasting error. Overall, it is interesting to see that traffic data are allowing us to create fairly precise estimates of GDP growth well before the official publication by Statistics Finland. Given the potentially real-time availability of traffic volumes' measurements, these results indicate the need to further explore the nowcasting ability of models based on these data.

The quarterly results reported in this subsection highlight the ability of models based on firm-level data and traffic data to provide accurate estimates of GDP growth. Even if the very early estimates, the ones computed during the quarter of reference, exhibit substantial nowcasting errors, the performance of our framework becomes significantly better when we consider the predictions at $t + 16$. While these flash estimates occur after the end of the quarter of reference, they allow for a 45 days reduction in the publication lag, which represents a substantial improvement.

6 Conclusions

We have examined the potential of large micro-level datasets, in combination with statistical models and machine learning techniques that are able to handle high-dimensional information sets, for the production of faster estimates of real economic activity indicators, both at the monthly and at the quarterly frequency. In particular, we have examined the nowcasting performance of firm-level data, and of trucks' traffic volumes measurements.

We find that a simple combination of the nowcasts obtained from a large set of machine learning techniques and large dimensional statistical models is able to produce accurate estimates of monthly real economic activity, or at least estimates that do not lead to a much larger revision error compared to the current official publications. While the revision errors do not increase substantially, our approach based on firm-level data allows for a reduction in the publication lag of roughly 30 days, when considering the monthly indicator. Turning to the results related to quarterly GDP, we find that our nowcasts would produce fairly accurate estimates of GDP growth during the third months of the reference quarter, even though there are few large errors. On the other hand, the nowcasts computed at $t + 16$ are accurate and do not show large revisions, or at least revisions that are compatible with the ones of Statistics Finland. Even though these estimates would be produced after the end of the quarter, they would still allow for more than a month reduction of the publication lag. Finally, it is important to

underline the satisfactory performance of traffic measurements data. The potential of this source of information should be explored further, given its real-time availability.

In the Finnish setting, the traffic loop data is open to the general public, while the firm level data is collected for the purpose of official statistics production and protected by the strict confidentiality standards of the statistical office. However, similar data collections exist in the other statistical offices of most countries, making our proposed approach and data source an interesting possibility for data users who need timely information on the state of the economy. Statistical offices have the possibility to increase their own relevance as information producers by using this kind of novel techniques. The relatively small investments that are required are related to modeling skills (in maintaining and updating the models) and adding a few features in the existing IT systems for storing information on the models, results and source data. The users of these types of estimates should be regularly informed about the expected and realized nowcast errors and revisions in the target indicators.

References

Knut Are Aastveit and Tørres Trovik. Estimating the output gap in real time: A factor model approach. *The Quarterly Review of Economics and Finance*, 54(2):180–193, 2014. doi: 10.1016/j.qref.2013.09.00.

Filippo Altissimo, Riccardo Cristadoro, Mario Forni, Marco Lippi, and Giovanni Veronese. New Eurocoin: Tracking Economic Growth in Real Time. *The Review of Economics and Statistics*, 92(4):1024–1034, November 2010.

S. Boragan 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.

Jushan Bai and Serena Ng. Boosting diffusion indices. *Journal of Applied Econometrics*, 24(4):607–629, 2009. doi: 10.1002/jae.1063.

Emanuele Baldacci, Dario Buono, George Kapetanios, Stephan Krische, Massimiliano

Marcellino, Gian Luigi Mazzi, and Fotis Papailias. Big Data and Macroeconomic Nowcasting: from data access to modelling . Technical report, December 2016.

Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1):188–205, September 2011.

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.

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The Generalized Dynamic-Factor Model: Identification And Estimation. *The Review of Economics and Statistics*, 82(4):540–554, November 2000.

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.

John Geweke. *The Dynamic Factor Analysis of Economic Time Series*. Latent Variables in Socio-Economic Models. 1977.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55 (4):665–676, May 2008.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

Rob Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software, Articles*, 27(3):1–22, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i03.

Julie Josse and François Husson. missmda: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software, Articles*, 70(1):1–31, 2016. ISSN 1548-7660.

Troy D. Matheson, James Mitchell, and Brian Silverstone. Nowcasting and predicting data revisions using panel survey data. *Journal of Forecasting*, 29(3):313–330, 2010. doi: 10.1002/for.1127.

Michele Modugno. Now-casting inflation using high frequency data. *International Journal of Forecasting*, 29(4):664–675, 2013. doi: 10.1016/j.ijforecast.2012.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

Thomas J. Sargent and Christopher A. Sims. Business cycle modeling without pretending to have too much a priori economic theory. Technical report, 1977.

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.

James H. Stock and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430, 2004. ISSN 1099-131X. doi: 10.1002/for.928. URL `http://dx.doi.org/10.1002/for.928`.

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

Klaus Wohlrabe and Teresa Buchen. Assessing the Macroeconomic Forecasting Performance of Boosting: Evidence for the United States, the Euro Area and Germany. *Journal of Forecasting*, 33(4):231–242, 07 2014.

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.