International Journal of Forecasting 29 (2013) 664–675

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Now-casting inflation using high frequency data

Michele Modugno

Université libre de Bruxelles, Belgium

ARTICLE INFO

Keywords:
Factor models
Forecasting
Inflation
Mixed frequencies

ABSTRACT

This paper proposes a methodology for now-casting and forecasting inflation using data with a sampling frequency which is higher than monthly. The data are modeled as a trading day frequency factor model, with missing observations in a state space representation. For the estimation we adopt the methodology proposed by Bańbura and Modugno (2010). In contrast to other existing approaches, the methodology used in this paper has the advantage of modeling all data within a single unified framework which allows one to disentangle the model-based news from each data release and subsequently to assess its impact on the forecast revision. The results show that the inclusion of high frequency data on energy and raw material prices in our data set contributes considerably to the gradual improvement of the model performance. As long as these data sources are included in our data set, the inclusion of financial variables does not make any considerable improvement to the now-casting accuracy.

1. Introduction

Forecasting inflation is of interest to both market practitioners and central banks. Market practitioners monitor macroeconomic releases continuously, among them inflation, in order to update their expectations on the future developments of macroeconomic fundamentals, and thus adjust their investment strategies. Central banks are charged with the task of guaranteeing price stability, and therefore routinely monitor inflation expectations and update inflation forecasts in their efforts to pinpoint the current inflation developments and to understand the underlying forces that may jeopardize price stability. In this context, now-casting inflation can help policy makers to make more accurate and timely monetary policy decisions.

This paper provides an econometric framework which allows interested parties to update their inflation forecasts continuously, following growing amounts of incoming information on a wide range of relevant available data series. It also allows the effects of different data groups on forecast revisions to be disentangled. The relevant data groups include variables which are highly correlated with inflation and which are released earlier than the relevant inflation releases. Moreover, they include data which are sampled at different frequencies, which allows information to enter the model in a sequential manner, as information updates become available.

The empirical analysis focuses on US CPI inflation. US CPI data are typically released at the middle of the month following the reference month. However, a wide range of data are released during the reference month at weekly or daily sampling frequencies, which carry valuable information on consumer prices. This information and the early signals that it contains can be useful for improving the accuracy of the predicted actual inflation and its future developments.

This approach was inspired by the relevant GDP now-casting literature. Indeed, although now-casting inflation is a novel idea, there is a rather long history of studies focusing on now-casting GDP. GDP is a quarterly variable released with a substantial time delay (e.g., one month after the end of the reference quarter for the US GDP), and several monthly indicators are released in the meantime. The seminal paper by Giannone, Reichlin, and Small (2008) for the US, and its first applications for the euro area, by Angelini, Bańbura, and Rünstler (2007) and Bańbura and

E-mail address: mmodugno@ulb.ac.be.

0169-2070/$ – see front matter © 2013 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2012.12.003

Rünstler (2007), show that using monthly indicators is crucial to accurate now-casting of the GDP.

Turning our focus to forecasting inflation, the topic has attracted the attention of various academics in the past. Attempts to forecast inflation include, among others, Cecchetti, Chu, and Steindel (2000) (who use univariate leading indicators models), Stock and Watson (1999) (who adopt factor models), Stock and Watson (2003) (who exploit asset prices in regression models) and Banerjee and Marcellino (2006) (who assess the relative forecast accuracy among an automated model selection procedure, a factor model, and single-indicator-based forecast pooling). However, these papers do not take into account the real-time availability and the mixed frequency nature of the data being used. Moreover, they focus mainly on long- rather than short-horizon predictions.

More recently, two different approaches have been proposed for the use of high frequency indicators in now-casting/forecasting inflation. In the first one, Lenza and Warmedinger (2011) use monthly, weekly and daily data for now-casting/forecasting inflation with a factor model. In order to add weekly and daily data to the monthly data, they use univariate autoregressive models to produce forecasts of the higher frequency data in order to fill in the missing data within each month. Once these forecasts have been produced, the available information and the forecasts produced are aggregated for each month, to obtain the monthly counterparts of the daily and weekly data. Then, a factor model is employed in order to produce predictions of inflation. In the second approach, a new generation of models, the Mixed Data Sampling Regression Models (MIDAS), originally proposed by Ghysels, Santa-Clara, and Valkanov (2004), have been used by Monteforte and Moretti (2010) to forecast inflation using a two-step approach. They extract principal components from a large sample of daily financial variables, and then use them in the forecasting equation for the target variable. In order to prevent over-parametrization, the MIDAS approach assumes that the response to the high frequency explanatory variables follows a distributed lag polynomial.

In contrast to the above-mentioned procedures, the methodology used in this paper models data with mixed frequencies within a single unified, coherent framework. The data are modelled as a trading day frequency factor model with missing observations, cast in a state space representation. The estimation adopts the methodology proposed by Bańbura and Modugno (2010), based on the EM algorithm for unobserved components models originally presented by Watson and Engle (1983). Doz, Giannone, and Reichlin (2006) later showed that this algorithm allows the estimation of factor models by maximum likelihood using large panels of data. Taking this into account, Bańbura and Modugno (2010) generalize Watson and Engle's (1983) methodology in order to deal with large panels which are characterized by arbitrary patterns of missing data.

In this manner, the econometric methodology proposed in this paper fully exploits the co-movement of data with different frequencies, and stresses the importance of timeliness. That is, our methodology allows the information content of each added variable to be preserved, as it avoids the need to introduce model-based forecasts for missing data and aggregation procedures which combine observed and forecasted data. In practice, this method also allows us to assess the importance of timeliness in the information that enters in the models when forecasting inflation, which is typically more pronounced in higher frequency data. This is achieved by comparing the forecast performance of a factor model which uses only monthly data with the forecast performance of the proposed factor model, which also uses weekly and daily data. In this paper, weekly data include Weekly Retail Gasoline and Diesel Prices (WRGDP), and daily data include the World Market Prices of Raw Materials (RMP) and some of the most important financial variables.

Finally, the proposed methodology also allows the unpredicted information content of the releases (news) to be linked with the forecast revisions of the target variable directly. Specifically, the proposed framework offers the possibility of forecasting all of the variables involved individually (unlike the MIDAS procedure, where the higher frequency variables are considered exogenous). Model-based news from each variable introduced can then be defined as the difference between the data released and the respective model predictions. We can thus disentangle the differing impacts of various releases on the forecast revision, which helps to increase our understanding of the variables which affect the forecast outcome. This is an additional feature that this paper aims to introduce for the now-casting of inflation.

The empirical evidence shows that higher frequency data, which are more timely than lower frequency ones, are necessary in order to produce more accurate inflation forecasts. The model suggested by this paper significantly outperforms (at short horizons, i.e., zero and one month(s) ahead) the model using lower frequencies only, as well as standard benchmark models for forecasting inflation (the Random Walk and Integrated Moving Average models).

Moreover, the empirical evidence evaluates the subcomponents of inflation as well as the importance of the different data groups in forecasting inflation. Looking at the forecasting performances of the subcomponents of inflation, our results suggest that the forecast improvement is due, to a larger extent, to the more accurate forecasts of the energy component and the transport component. Furthermore, the inclusion of RMP and WRGDP improves the accuracy of the model significantly. However, as in Stock and Watson (2003), the evidence is not equally supportive of financial data, which appear to have only a marginal contribution, when utilized together with the other variable groups.

The paper is organized as follows: Section 2 describes the data, and Section 3 introduces the model and the estimation methodology. Section 4 presents the results from the forecast exercise, after which Section 5 explains the concept of news and presents an illustrative example. Finally, Section 6 concludes.

2. Data

The data included in the analysis have two main characteristics: a high correlation with inflation, and timeliness,

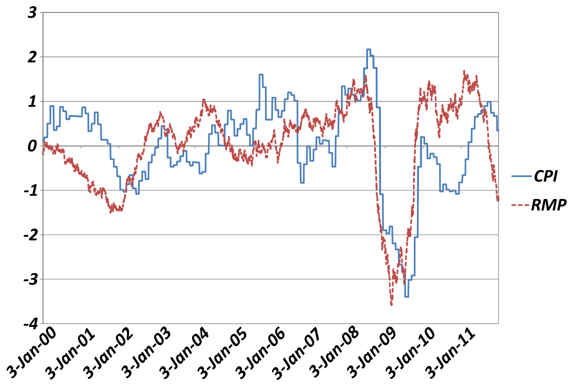

Fig. 1. Raw material prices. *Notes*: This figure shows the evolution of the year-on-year growth rates of overall CPI, and of the average of the eleven raw material price (RMP) series. All of the series are standardized and centered on zero.

i.e. an earlier release than the inflation data. These characteristics can help to improve the accuracy of inflation forecasts. In practice, the data set contains four groups of variables, plus the total CPI and its main components (energy, services and goods, transport, food and beverages, housing). Three of the groups relate to consumer prices directly, and capture raw material, energy and manufacturing and non-manufacturing prices at various stages of the pricing chain. The remaining group includes financial data, which aim to capture the general economic conditions that can eventually affect consumer prices. A list of all of the data is provided in Table 4 in Appendix.

More precisely, the first group of variables in our data set includes the World Market Price of Raw Materials (RMP). These data are sampled at a daily frequency and are published weekly on the second or third day of the week following the reference week. They are produced by the OECD as weighted averages of the commodity imports of OECD countries. This group of variables may capture global price dynamics, as well as pricing information in the early part of the pricing chain. They can therefore provide an indication of fundamental price developments. Fig. 1 shows that there is indeed a high degree of correlation between the average year-on-year growth rate of RMP series and the year-on-year growth rate of total CPI. This is a desirable feature for data which are employed for now-casting inflation.

The second group contains energy prices for fuel in the US. On Mondays, the Energy Information Administration collects and publishes the Weekly Retail Gasoline and Diesel Prices (WRGDP), a survey of the pump prices for gasoline and diesel. These data include taxes, and reflect the prices paid by the consumers. Compared to raw oil prices, they have the advantage that distribution and retail margins are fully accounted for. Therefore, the information they contain captures later stages of the pricing chain, and is more akin to consumer price information. Similarly to Fig. 1, Fig. 2 reveals that the average year-on-year growth rate of the WRGDP series is highly correlated with the year-on-year growth rate of total CPI.

The third group includes the ISM (Institute for Supply Management) manufacturing and non-manufacturing price indexes, which are monthly surveys which become available on the first working day of the month following the reference period, i.e. two weeks prior to the CPI data. Therefore, this data group is released in a more timely fashion than inflation and contains information on the last stage of the pricing chain for a broad range of variables (both manufacturing and non-manufacturing, i.e. services). Looking at Fig. 3, these two surveys co-move to a large extent with the year-on-year growth rate of total CPI.

The fourth group includes financial data at a daily frequency. In particular, it includes the trade-weighted US dollar index (major currencies), the *S&P* 500 index, the ten-year Treasury constant maturity rate, and the three-month Treasury-bill rate. Such financial variables may incorporate expectations on the future development of prices, and are available on daily basis; they therefore have the potential to be informative about the future development of our target variable.

Figs. 4–7 plot the chosen financial variables (the year-on-year growth rate of the *S&P* 500 index and the yearly differences for the other variables) *vis-à-vis* the year-on-year total CPI inflation. As we can see from Fig. 4, the yearly exchange rate differences appear to lag behind the total CPI inflation. In fact, Fig. 5 shows that the year-on-year growth rate of the *S&P* 500 index appears to have leading indicator

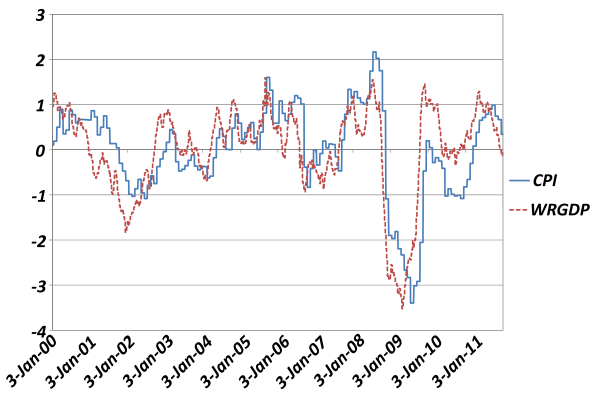

Fig. 2. Weekly retail gasoline and diesel prices.
Notes: This figure shows the evolution of the year-on-year growth rates of overall CPI, and of the average of the four series reported in the weekly retail gasoline and diesel prices (WRGDP). All of the series are standardized and centered on zero.

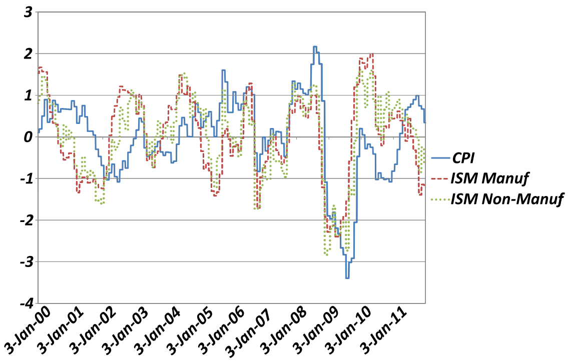

Fig. 3. Institute for Supply Management manufacturing and non-manufacturing prices indexes. *Notes*: This figure shows the evolution of the year-on-year growth rates of overall CPI, and the Institute for Supply Management manufacturing (ISM Manuf) and non-manufacturing (ISM Non-Manuf) price indexes. All of the series are standardized and centered on zero.

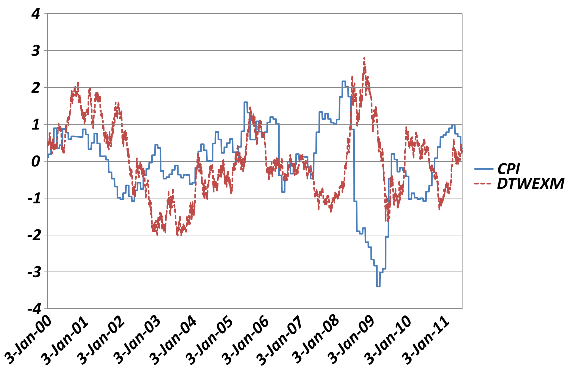

Fig. 4. Trade weighted US dollar index: major currencies. *Notes*: This figure shows the evolution of the year-on-year growth rates of overall CPI, and the yearly differences of the trade weighted US dollar index (DTWEXM). All of the series are standardized and centered on zero.

properties *vis-à-vis* total CPI inflation. With regard to the ten-year Treasury constant maturity rate and the three-month Treasury bill rate, shown in Figs. 6 and 7 as yearly differences, while the former does not appear to have a strong contemporaneous correlation with the total CPI, the latter seems to be more closely related to the low frequency behavior of the total CPI inflation. Overall, the evidence on the correlation of the financial data with total inflation does not point to a consistent direction among the various variables. However, we include these data as they are the timeliest data in the current data set.

Finally, because this paper focuses on the forecasts of CPI (total and components), the data set includes six CPI series, i.e. the total CPI plus its components: energy, food and beverages, housing, goods and services, and transport. By assessing the effects of the remaining groups of variables (WRGDP, RMP, financial and ISM data) on the forecast of each individual CPI component, we shed light on the underlying forces driving an improvement in the forecast accuracy of the total CPI.

In order to enable a better and easier understanding of the timeliness of each data release with respect to the CPI data, we present a timeline of the various groups of variables employed for a given month as they appear. Overall, the following logic applies: the first release of inflation data in the US is published about 15 days after the reference month. Prior to that day, however, the remaining variables in our data set, which already contain preliminary evidence/information about the CPI of the current month, have already been released. Fig. 8 gives an illus-

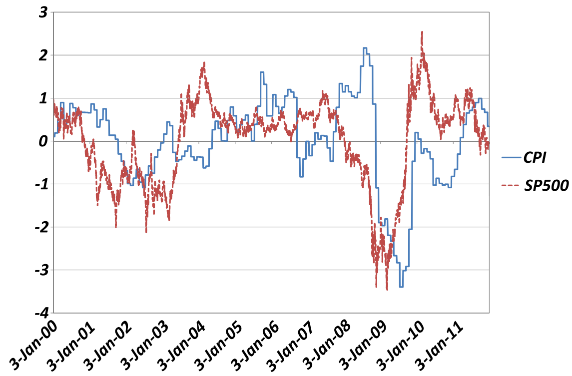

Fig. 5. *S&P* 500 index. *Notes*: This figure shows the evolution of the year-on-year growth rates of overall CPI, and of the Standard and Poor's 500 index (*S&P* 500). All of the series are standardized and centered on zero.

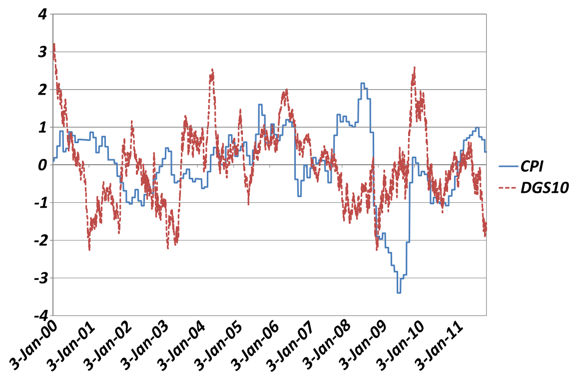

Fig. 6. Ten-year Treasury constant maturity rate. *Notes*: This figure shows the evolution of the year-on-year growth rates of overall CPI, and the yearly differences of the ten-year treasury constant maturity rate (DGS10). All of the series are standardized and centered on zero.

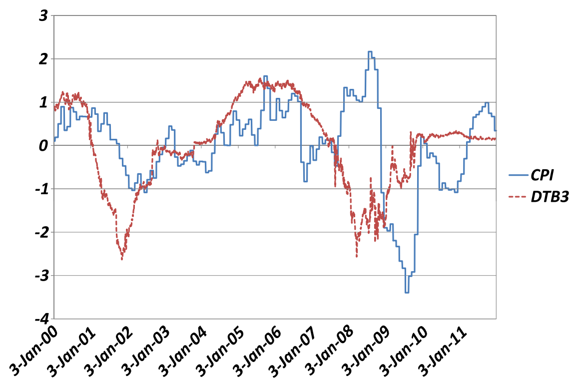

Fig. 7. Three-month Treasury bill rate. *Notes*: This figure shows the evolution of the year-on-year growth rates of overall CPI, and the yearly differences of the three-month Treasury bill rate (DTB3). All of the series are standardized and centered on zero.

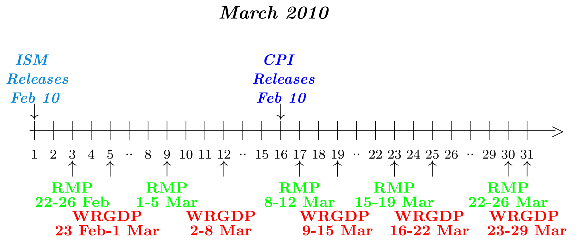

Fig. 8. Timeliness. *Notes*: This figure shows the flow of data released in a specific month, March 2010. These data are the weekly retail gasoline and diesel prices (WRGDP), the world market price of raw materials (RMP), the consumer prices indexes (CPI) and the Institute for Supply Management manufacturing and non-manufacturing prices indexes (ISM).

trative example of the timeliness of the data employed to forecast the CPI inflation in March 2010, our reference period in this example. It should be noted that the CPI data for March 2010 (i.e., the target forecast) are released on 15 April, and therefore are not presented in our March timeline. As far as CPI data releases in March are concerned, the only available information is the CPI for February, which is released on 16 March (see the upper part of Fig. 8).

Focusing on the four variable groups of our data set, it becomes clear that they are released in a much more timely manner than the CPI data. Starting at the beginning of the month, the monthly ISM surveys are released on March 1, and contain information about the previous month, i.e., February 2010 (see the upper part of Fig. 8). Now let us consider the lower part of Fig. 8, which shows the arrival of the higher frequency information in March 2010. During this reference month, several sets of WRGDP and RMP data are released, containing information about February and (mainly) March. Specifically, the first two releases of RMP and WRGDP, on March 3 and March 5 respectively, contain information about February (the RMP data relate to the trading days from February 22 to February 26, and the WRGDP data relate to the week from February 23 to March 1). The information flow for March begins on March 9, with the RMP for the period from March 1 to March 5. This information is updated on a weekly basis (March 17, 23 and 30). In a parallel fashion, the WRGDP data are released starting from 12 March (for the first week of March), and are also updated at weekly intervals (March 19, 25 and 31). Moreover, financial data for the current day are also available on each trading day (not shown in the timeline).

3. Methodology

This paper adopts a dynamic factor model in order to produce forecasts of inflation. Factor models avoid overparametrization by summarizing all of the data employed in few unobserved components, which capture the comovement among the data. This allows us to exploit the information contained in many series contemporaneously.

3.1. Estimation

The general representation of a dynamic factor model is:

$$y_t = Cf_t + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d.}\, N(0, \Sigma), \tag{1}$$

where y_t is an $n \times 1$ vector of observations, C is an $n \times r$ matrix of loadings, ϵ_t is an $n \times 1$ vector of series specific idiosyncratic components, and f_t is an $r \times 1$ vector of unobserved common components that display VAR dynamics:

$$f_t = A(L)f_{t-1} + u_t, \quad u_t \sim \text{i.i.d.}\, N(0, Q), \tag{2}$$

where u_t is an $r \times 1$ vector, and A is an $r \times r$ matrix.

From now on, t will refer to trading days. The general representation of a dynamic factor model can be estimated in several different ways. In this specific case, we have to choose an estimation methodology that can deal with two issues. First, given that the vector y_t consists of data observed at a range of different frequencies, we observe missing data when modeling at the highest frequency available (in our case, daily). Following Mariano and Murasawa (2003), we indeed assume that lower frequency variables are missing periodically. Second, as we will show in Section 3.2, we will need to impose restrictions on the parameters C, A, Q and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$.

In order to deal with these two issues, we adopt the estimation methodology proposed by Bańbura and Modugno (2010), which generalizes the methodology proposed by Watson and Engle (1983) to the case of missing data. The latter methodology is based on the Expectation Maximization (EM) algorithm under the assumption of an exact factor model, i.e., without serial and cross-correlation in the idiosyncratic components. However, Doz et al. (2006) argue that these assumptions could be too restrictive. In particular, for the case of large cross-sections, they study the approximate factor model, allowing weak serial and cross-correlation in the idiosyncratic component. They show that as $n, T \to \infty$, the factors can be estimated consistently by quasi maximum likelihood, i.e., assuming that the model is a misspecification of the exact factor model (for the technical details, see Doz et al., 2006). Consequently, the estimators are asymptotically valid in the case of approximate factor models as well. Bańbura and Modugno (2010) also show how the parameters should be estimated in the case of an arbitrary pattern of missing data, e.g., when the data set includes data which are sampled at different frequencies or with varying publication lags. They tackle this issue by deriving the parameters C, A, Q and Σ under the assumption that the observations are defined as follows:

$$y_t = W_t y_t + (I_n - W_t)y_t, \tag{3}$$

where W_t is a diagonal matrix of size n, where the ith diagonal element is equal to 0 if y_{it} is missing and equal to 1 otherwise; I_n is an identity matrix of dimension. This allows a factor structure to be imposed on the ith variable only when y_{it} is available. Moreover, Bańbura and Modugno (2010) show how restrictions can be imposed on the parameters, in order to impose a block structure on the factor model.

3.2. Econometric framework

As was explained in Section 2, the n variables composing our data set are divided into three groups with different sampling frequencies: monthly (m), weekly (w) and daily (d). Let us define:

- $Y_t^{(m)}$ as the logarithm of the monthly series Y in month m and on day t. There are k_m trading days between any two consecutive releases, where this value can vary from 15 to 23, depending on the month m.
- $Y_t^{(w)}$ as the logarithm of the weekly series Y on week w and day t. There are k_w trading days between any two consecutive releases, with this value usually being 5, as long as there are no bank holidays in week w.
- $Y_t^{(d)}$ as the logarithm of the daily series Y, observed on each day t.[1]

Given these definitions for the log-levels, we derive:

- $y_t^{(m)} = (Y_t^{(m)} - Y_{t-k_m}^{(m)}) \times 100$, i.e., the monthly growth rate for the series sampled at a monthly frequency.
- $y_t^{(w)} = (Y_t^{(w)} - Y_{t-k_w}^{(w)}) \times 100$, i.e., the weekly growth rate for the series sampled at a weekly frequency.
- $y_t^{(d)} = (Y_t^{(d)} - Y_{t-1}^{(d)}) \times 100$, i.e., the daily growth rate for the series sampled at a daily frequency.[2]

Our vector of observables is then $y_t = [y_t^{(m)}, y_t^{(w)}, y_t^{(d)}]'$.

Given that the three groups of variables express three different measures, i.e. monthly, weekly and daily growth rates, we have to modify Eqs. (1) and (2) in order to have a coherent model, where the unobserved components are expressed using the same measure of the specific series that they load. We chose the following representation:

$$\begin{bmatrix} y_t^{(m)} \\ y_t^{(w)} \\ y_t^{(d)} \end{bmatrix} = \begin{bmatrix} C_m & 0 & 0 \\ 0 & C_w & 0 \\ 0 & 0 & C_d \end{bmatrix} \begin{bmatrix} f_t^{(m)} \\ f_t^{(w)} \\ f_t^{(d)} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^{(m)} \\ \varepsilon_t^{(w)} \\ \varepsilon_t^{(d)} \end{bmatrix}, \tag{4}$$

where the transition equation becomes:

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_t^{(m)} \\ f_t^{(w)} \\ f_t^{(d)} \end{bmatrix} = \begin{bmatrix} \Xi_t^{(m)} & 0 & 0 \\ 0 & \Xi_t^{(w)} & 0 \\ 0 & 0 & A \end{bmatrix} \begin{bmatrix} f_{(t-1)}^{(m)} \\ f_{(t-1)}^{(w)} \\ f_{(t-1)}^{(d)} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ u_t^{(d)} \end{bmatrix}, \tag{5}$$

where C_m, C_w and C_d are the loadings for monthly, weekly and daily variables, respectively; $f_t^{(m)}$, $f_t^{(w)}$ and $f_t^{(d)}$ are the monthly, weekly and daily factors; Ξ_t^w is a time varying coefficient which is equal to zero the day after each release of the weekly data and to one elsewhere; Ξ_t^m is equal to zero the day after each release of the monthly data and to one elsewhere; and A is the matrix of the autoregressive coefficients for the daily factors (for illustrative purposes, we assume that $f_t^{(d)}$ is characterized by a VAR(1) dynamic). Once the model has been written in this state space form, it is straightforward to apply the methodology proposed by Bańbura and Modugno (2010).

A more detailed view of the state space representation of the model may be informative. As was described in Section 2, CPI indexes are collected around the 15th of each month. As such, they can be considered as a snapshot of the prices around that day. In order to explain how this paper models snapshot variables observed at a monthly frequency within a daily factor model, suppose that the monthly variables were sampled every day, and take their daily growth rate, $\hat{y}_t^{(m)} = (Y_t^{(m)} - Y_{t-1}^{(m)}) \times 100$. The monthly growth rate can be derived from the daily one by summing the daily growth rates from the first day after the previous month's release to the day of the actual release:

$$\begin{aligned} y_t^{(m)} &= \sum_{i=t-k_m+1}^{t} \hat{y}_i^{(m)} = (Y_t^{(m)} - Y_{t-1}^{(m)} + Y_{t-1}^{(m)} - Y_{t-2}^{(m)} \\ &\quad + Y_{t-k_m+1}^{(m)} - Y_{t-k_m}^{(m)}) \times 100 \\ &= (Y_t^{(m)} - Y_{t-k_m}^{(m)}) \times 100. \end{aligned}$$

This implies that the monthly growth rate of a variable can be recovered by summing its daily growth rates, if available. There are no daily growth rates for CPI data, but daily data that co-move with CPI can be used to extract a common factor at a daily frequency, $f_t^{(d)}$. Given the availability of $f_t^{(d)}$, $y_t^{(m)}$ can be rewritten as:

$$y_t^{(m)} = \sum_{i=t-k_m+1}^{t} \hat{y}_i^{(m)} = C_m \sum_{i=t-k_m+1}^{t} f_i^{(d)} = C_m f_t^{(m)},$$

where $f_t^{(m)}$ is the sum of the daily factors ($f_t^{(d)}$) from day $t - k_m + 1$ to day t.

In order to aggregate the daily factors – and keeping in mind that Ξ_t^m is a time varying coefficient that is equal to zero the day after each release of the monthly data and to one elsewhere – we can define:

$$f_t^{(m)} = \Xi_t^m f_{t-1}^{(m)} + f_t^{(d)}. \tag{6}$$

This is the definition of $f_t^{(m)}$ expressed in Eq. (5). It means that on the first day after the previous month's release, and given that, by definition, $\Xi_{t-k_m}^m = 0$ and $\Xi_{t-k_m+1}^m = 1$, Eq. (6) implies that $f_{t-k_m+1}^{(m)} = f_{t-k_m+1}^{(d)}$. On the second day after the previous month's release, again by definition $\Xi_{t-k_m+2}^m = 1$, and consequently $f_{t-k_m+2}^{(m)} = \Xi_{t-k_m+2}^m f_{t-k_m+1}^{(m)} + f_{t-k_m+2}^{(d)} = f_{t-k_m+1}^{(d)} + f_{t-k_m+2}^{(d)}$. Iterating this sum for the remaining days between two consecutive releases of monthly data, on the day of the release for the current month (t), we have that $f_t^{(m)} = \Xi_t^m f_{t-1}^{(m)} + f_t^{(d)} = \sum_{i=t-k_m+1}^{t} f_i^{(d)}$. The same explanation applies for weekly data.

3.3. Benchmark models

Throughout the empirical section, we will compare the performances of the factor models with those of two benchmark models, a random walk without drift and a first-order integrated moving average process, from now on referred to as IMA(1, 1), as proposed by Stock and Watson (2007). Denoting a CPI inflation series by π_t, the RW model for this series is:

$$\pi_t = \pi_{t-1} + \epsilon_t, \tag{7}$$

[1] In this case, (d) is used for illustrative purposes and in analogy with (m) and (w), but it essentially refers to the same trading day t.

[2] The financial variables, with the exception of the S&P 500 index, are included as daily differences.

where ϵ_t is serially uncorrelated, with mean zero and variance σ_ϵ^2.

Instead, the IMA(1, 1) process for π_t is:

$$\pi_t = \pi_{t-1} + a_t - \beta a_{t-1}, \quad (8)$$

where a_t is serially uncorrelated, with mean zero and variance σ_a^2.

4. Forecast exercise: design and results

In this section, we begin by explaining the design of the forecasting exercise. We then present the competing models and describe the process chosen for evaluating the forecast performances. Finally, we present the results.

4.1. Forecast exercise design

In order to determine whether high frequency data can improve the inflation forecast accuracy, we compare the following models:

- a factor model that includes only monthly variables (*Mon*), estimated at a monthly frequency, i.e., CPI indexes and ISM surveys;
- a factor model that includes all of the variables (*All*), estimated at the trading day frequency, as explained in Section 3.2;
- a naïve random walk model (RW);
- an IMA(1, 1) model.

We adopt a recursive estimation scheme which, for the first evaluation, covers the period from April 1996 to January 2000. The evaluation sample spans the period from January 2001 to December 2011. We evaluate the forecasts at 0 (now-cast), 1, 3, 6 and 12 months ahead, and the results are expressed as root mean squared forecast errors (RMSFE).

The data under analysis are never revised (see Giannone, Henry, Lalik, & Modugno, 2012); this implies that, taking into account the data availability at each point in time at which we produce forecasts, we perform a real-time forecasting exercise.

Fig. 8 helps to illustrate the exercise better. There, we see that the information enters into the estimation gradually. The quality of the information content of each release affects the accuracy of the now-cast/forecast. Our model can be evaluated on any given day of the month. In this section, we show the results obtained the day after CPI data for the previous month are released. For example, in March 2010 (see Fig. 8), we produce the forecasts on the 17th of March, the day after the CPI data for February are released. At that point in time, we already have three releases of higher frequency data that contain information about March: two for RMP data and one for WRGDP data, available on March 9 and 17 and March 12, respectively. On the other hand, at that point in time, we do not have any CPI data for the current month, and the information available at a monthly frequency is the CPI and survey releases for the previous month (February). Therefore, on the 17th, the only data available for the month that we are now-casting (in this example, March 2010) are WRGDP, RMP and financial data. In Section 5, we show the evolution of the forecast accuracy obtained when evaluating the models at various different points in time.

4.2. Results

This section presents the results of the alternative models described in Section 4.1. Results are presented for the model that produces the best now-cast ex-post (one factor and six lags in the transition equation). However, in order to mimic a proper out-of-sample forecasting exercise, we also present the arithmetic average of the RMSFEs produced by 24 factor models which are characterized by different parameterizations (one or two factors and one to twelve lags).[3] The first model uses all data available (*All*), while the second uses only monthly data (*Mon*). Moreover, we also report the results obtained with two univariate models, namely a Random Walk (RW) and an IMA(1, 1), as proposed by Stock and Watson (2007).

The results for the full sample are reported in Table 1. The table shows that the inclusion of WRGDP, RMP and financial data improves the forecast accuracy for the total CPI at all horizons (with this improvement being statistically significant for forecasts up to 3 months). This appears to be particularly true for the now-cast (horizon 0) of total CPI, as the RMSFE for model *All* is 0.23 at horizon 0, while the RMSFE for model *Mon* for the same horizon is 0.33. Most of this gain is due to the better forecasting performance of model *All* for the energy and transport components. Indeed, the RMSFE for the now-cast of the energy component is 2.28 for model *All*, while it is 3.53 for model *Mon*. For the transport component, the RMSFE produced by model *All* at horizon 0 is 1.12, while model *Mon* produces a RMSFE of 1.72. For the remaining CPI components, the results suggest that the inclusion of daily and weekly data in the data set does not improve the forecast accuracy significantly. Indeed, for the food and beverages component, models *All* and *Mon* produce equally accurate forecasts. With regard to the housing and goods and services components, model *All* performs slightly worse than model *Mon*. This could be due to the lack of timely, higher frequency data for the housing and service sectors. With regard to the benchmark models, i.e., the Random Walk and the IMA(1, 1), they are consistently out-performed by the forecasts produced by all of the factor models at all horizons for the total CPI, as well as for most of the subcomponents. Moreover, the model with the lowest RMSFE, i.e. the best model ex-post (*BXP*), differs only marginally (at the third decimal place) from the arithmetic average of the RMSFEs produced by 24 factor models characterized by different parameterizations. Therefore, the parametrization does not affect the robustness of the results.

[3] One of the most hotly debated topics in the factor model literature is how to choose the number of factors in Eq. (1) and the number of lags for the VAR in Eq. (2), especially for forecasting purposes. Several solutions have been proposed, but there has been no clear consensus. This is the reason why, in order to mimic a proper out-of-sample forecasting exercise, this papers presents results for the model that produces the best now-cast ex-post (one factor and six lags in the transition equation), and also the arithmetic average of the RMSFEs produced by 24 factor models which are characterized by different parameterizations.

Table 1
Forecast evaluation: full sample.

Horizon	*All*	*Mon*	IMA	RW	*BXP*	*All*	*Mon*	IMA	RW	*BXP*	*All*	*Mon*	IMA	RW	*BXP*
	Total					Energy					Food and beverages				
12	1.46	1.49	1.71	2.09	1.45	15.32	15.10	18.90	21.92	15.25	1.57	1.56	1.93	2.04	1.57
6	1.11	1.18	1.49	1.65	1.10	11.77	12.17	15.44	17.32	11.71	1.02	1.02	1.36	1.38	1.02
3	0.81*	0.93	1.22	1.28	0.80	8.37	9.53	12.13	13.00	8.31	0.64	0.65	0.89	0.89	0.64
1	0.45***	0.62	0.86	0.87	0.44	4.44***	6.52	8.88	9.11	4.32	0.37	0.37	0.51	0.51	0.37
0	0.23***	0.33	0.52	0.52	0.23	2.28***	3.53	5.50	5.54	2.27	0.22	0.22	0.31	0.30	0.22
	Housing					Goods and services					Transport				
12	1.61	1.62	2.13	1.45	1.61	2.60	2.30	4.83	2.32	2.61	7.01	6.99	8.44	10.48	6.96
6	0.93	0.94	1.17	0.97	0.93	1.62	1.45	1.58	1.64	1.62	5.66	5.92	7.37	8.55	5.61
3	0.57	0.58	0.73	0.66	0.57	1.12	1.02	1.29	1.32	1.12	4.11*	4.68	5.99	6.54	4.06
1	0.32	0.33	0.44	0.41	0.32	0.69	0.65	0.94	0.94	0.69	2.26***	3.17	4.32	4.48	2.19
0	0.19	0.19	0.26	0.25	0.19	0.50	0.48	0.61	0.60	0.50	1.12***	1.72	2.66	2.68	1.11

Notes: This table shows the root mean squared forecast errors (RMSFEs) produced by the factor models, the random walk and the IMA(1, 1) for overall CPI inflation and its components. *All* and *Mon* are averages of the RMSFEs produced by factor models with 1 or 2 factors in Eq. (1), and from 1 to 12 lags in Eq. (2). The two factor models differ in terms of the data sets used. *Mon* uses only the monthly data listed in Table 4. *All* uses all of the available data. *BXP* reports the RMSFEs produced by the best model ex-post, i.e., the factor model that uses all of the available data with one factor and six lags.

* Indicates that the null hypothesis of the Diebold and Mariano (1995) test for equal accuracy of model *All* versus model *Mon* is rejected with a *p*-value equal to or smaller than 10%.

*** Indicates that the null hypothesis of the Diebold and Mariano (1995) test for equal accuracy of model *All* versus model *Mon* is rejected with a *p*-value equal to or smaller than 1%.

Table 2
Forecast evaluation: sub-sample January 2001–December 2005.

Horizon	*All*	*Mon*	IMA	RW	*BXP*	*All*	*Mon*	IMA	RW	*BXP*	*All*	*Mon*	IMA	RW	*BXP*
	Total					Energy					Food and beverages				
12	0.97	0.90	1.43	1.18	0.95	14.25	11.45	19.49	16.28	14.15	0.91	0.83	1.02	1.32	0.85
6	0.68	0.65	0.81	0.88	0.68	9.94	8.63	11.20	12.12	10.01	0.58	0.55	0.78	0.87	0.54
3	0.56	0.54	0.67	0.70	0.57	7.64	7.02	8.91	9.24	7.80	0.40	0.38	0.56	0.58	0.38
1	0.36***	0.42	0.58	0.61	0.35	4.44***	5.27	7.41	7.62	4.34	0.28	0.27	0.37	0.38	0.27
0	0.21***	0.24	0.37	0.38	0.20	2.53***	3.02	4.71	4.76	2.43	0.19	0.19	0.27	0.27	0.19
	Housing					Goods and services					Transport				
12	0.93	0.95	2.36	1.17	0.94	3.22	2.44	1.85	1.31	3.16	4.56	4.07	7.48	5.45	4.47
6	0.60	0.58	1.12	0.80	0.60	1.85	1.32	1.28	0.88	1.78	3.24	3.01	3.87	4.04	3.23
3	0.42	0.42	0.64	0.56	0.42	1.20	0.87	0.89	0.66	1.15	2.78	2.56	3.17	3.29	2.83
1	0.29	0.28	0.39	0.37	0.29	0.67	0.50	0.58	0.46	0.64	1.81***	2.14	2.93	3.04	1.76
0	0.19	0.18	0.26	0.25	0.19	0.59	0.53	0.53	0.52	0.57	1.06***	1.27	1.90	1.93	1.03

Notes: This table shows the root mean squared forecast errors (RMSFEs) produced by the factor models, the random walk and the IMA(1, 1) for overall CPI inflation and its components. *All* and *Mon* are averages of the RMSFEs produced by factor models with 1 or 2 factors in Eq. (1), and from 1 to 12 lags in Eq. (2). The two factor models differ in terms of the data set used. *Mon* uses only the monthly data listed in Table 4. *All* uses all of the available data. *BXP* reports the RMSFEs produced by the best model ex-post, i.e., the factor model that uses all of the available data with one factor and six lags .

*** Indicates that the null hypothesis of the Diebold and Mariano (1995) test for equal accuracy of model *All* versus model *Mon* is rejected with a *p*-value equal to or smaller than 1%.

In Tables 2 and 3 we report the RMSFEs for two sub-samples, from January 2001 to December 2005 and from January 2006 to December 2011 respectively. As we can see, the results for the factor models are quite stable in the two sub-samples. In particular, while model *All* outperforms the competitors at horizons zero and one (i.e., the now-cast and 1-month-ahead forecast) for the total CPI and its energy and transport components on the first sub-sample, it performs well even for the 3-month-ahead forecast in the second sub-sample. These results indicate that the correlation between the high frequency data and the CPI inflation is rather strong in both sub-samples. Overall, it appears that the results on the full sample are not driven by the large fluctuations experienced in total CPI inflation, RMP and WRGDP over recent years.

5. *News* and forecast revisions: do financial variables matter?

Section 4 demonstrates the performances of our model, assuming that forecasts are produced once per month, i.e., after the CPI release. However, when forecasting in real-time, there is a continuous inflow of information as new figures for various predictors are released non-synchronously and with different degrees of delay. Therefore, in such applications, we seldom perform a single prediction for the reference period, but rather produce a sequence of forecasts, which are updated when new data arrive. Intuitively, only the news or unexpected components of the data released should revise the forecast; hence, extracting the news and linking it to the resulting forecast revision is a key part of understanding and interpreting the latter.

Table 3
Forecast evaluation: sub-sample January 2006–December 2011.

Horizon	*All*	*Mon*	IMA	RW	*BXP*	*All*	*Mon*	IMA	RW	*BXP*	*All*	*Mon*	IMA	RW	*BXP*
	Total					Energy					Food and beverages				
12	1.76	1.77	1.98	2.70	1.76	16.01	15.83	18.81	26.39	16.10	1.98	1.98	2.56	2.61	1.97
6	1.37	1.44	1.91	2.18	1.36	13.07	13.64	18.18	21.60	12.95	1.29	1.28	1.78	1.77	1.28
3	0.97*	1.12	1.59	1.69	0.95	8.97*	10.65	14.75	16.22	8.70	0.80	0.79	1.14	1.12	0.80
1	0.52***	0.71	1.06	1.09	0.50	4.57***	6.99	10.21	10.66	4.29	0.44	0.43	0.63	0.62	0.44
0	0.25***	0.37	0.63	0.63	0.24	2.15***	3.65	6.26	6.34	2.13	0.24	0.24	0.35	0.34	0.24
	Housing					Goods and services					Transport				
12	2.01	1.96	2.06	1.67	2.01	2.03	2.04	2.13	2.62	2.03	8.47	8.15	9.52	13.77	8.50
6	1.14	1.09	1.22	1.10	1.14	1.46	1.48	1.75	1.90	1.46	7.07	7.22	9.47	11.44	7.00
3	0.67	0.64	0.81	0.74	0.67	1.10	1.10	1.42	1.48	1.10	4.99**	5.71	7.84	8.71	4.85
1	0.35	0.33	0.47	0.44	0.35	0.74	0.73	1.01	1.02	0.74	2.63***	3.73	5.40	5.64	2.48
0	0.19	0.18	0.27	0.26	0.19	0.44	0.44	0.63	0.62	0.44	1.20***	1.97	3.27	3.30	1.17

Notes: This table shows the root mean squared forecast errors (RMSFEs) produced by the factor models, the random walk and the IMA(1, 1) for overall CPI inflation and its components. *All* and *Mon* are averages of the RMSFEs produced by factor models with 1 or 2 factors in Eq. (1), and from 1 to 12 lags in Eq. (2). The two factor models differ in terms of the data set used. *Mon* uses only the monthly data listed in Table 4. *All* uses all of the available data. *BXP* reports the RMSFEs produced by the best model ex-post, i.e., the factor model that uses all of the available data with one factor and six lags .

* Indicates that the null hypothesis of the Diebold and Mariano (1995) test for equal accuracy of model *All* versus model *Mon* is rejected with a *p*-value equal to or smaller than 10%.

** Indicates that the null hypothesis of the Diebold and Mariano (1995) test for equal accuracy of model *All* versus model *Mon* is rejected with a *p*-value equal to or smaller than 5%.

*** Indicates that the null hypothesis of the Diebold and Mariano (1995) test for equal accuracy of model *All* versus model *Mon* is rejected with a *p*-value equal to or smaller than 1%.

In contrast to models that consider high frequency data as predetermined regressors (e.g., MIDAS), the model employed in this paper has the advantage that it treats all of the variables as endogenous. This implies that our unique framework produces forecasts for all of the variables, and therefore it gives us the possibility of extracting a model-based unexpected component for each variable when new figures are released. This section begins by showing how the unexpected content, i.e., the *news*, in a data release is linked to the resulting forecast revision. It then describes how this news can be used to evaluate the importance of different groups of variables, in order to backcast, now-cast and forecast total CPI inflation.

5.1. News and forecast revision

Let Ω_{v-1} and Ω_v be two consecutive vintages of data, and consequently $\Omega_{v-1} \subset \Omega_v$.[4] Let I_v denote the *news* in Ω_v with respect to Ω_{v-1}. For example, let us assume that the difference between Ω_{v-1} and Ω_v is the release of RMP data for the period t_i. The *news* is $I_v = y_{t_i}^{\text{RMP}} - \mathrm{E}\left(y_{t_i}^{\text{RMP}}|\Omega_{v-1}\right)$, where $y_{t_i}^{\text{RMP}}$ is a vector containing the last figures released. Assume that we are interested in the way in which *news* revises the total CPI inflation forecast for the period t_j. As $I_v \perp \Omega_{v-1}$, we can write

$$\mathrm{E}\left(\cdot|\Omega_v\right) = \mathrm{E}\left(\cdot|\Omega_{v-1}\right) + \mathrm{E}\left(\cdot|I_v\right)$$

or

$$\underbrace{\mathrm{E}\left(y_{t_j}^{\text{CPI}}|\Omega_v\right)}_{\text{new forecast}} = \underbrace{\mathrm{E}\left(y_{t_j}^{\text{CPI}}|\Omega_{v-1}\right)}_{\text{old forecast}} + \mathrm{E}\Big(y_{t_j}^{\text{CPI}}|\underbrace{I_v}_{\text{news}}\Big).$$

In other words, the updated forecast can be decomposed into the sum of the old forecast and the contribution from the *news* in the latest release. To compute the latter, we use the fact that

$$\mathrm{E}\left(y_{t_j}^{\text{CPI}}|I_v\right) = \mathrm{E}\left(y_{t_j}^{\text{CPI}}I_v'\right)\mathrm{E}\left(I_vI_v'\right)^{-1}I_v.$$

Furthermore, given Eq. (1), we can write

$$y_{t_j}^{\text{CPI}} = C_{\text{CPI}}f_{t_j} + \epsilon_{t_j}^{\text{CPI}},$$

$$I_v = y_{t_i}^{\text{RMP}} - y_{t_i|\Omega_{v-1}}^{\text{RMP}} = C_{\text{RMP}}\left(f_{t_i} - f_{t_i|\Omega_{v-1}}\right) + \epsilon_{t_i}^{\text{RMP}},$$

where C_{CPI} and C_{RMP} are the rows of C corresponding to CPI and RMP, respectively. It can be shown (see Bańbura & Modugno, 2010) that:

$$\begin{aligned}\mathrm{E}\left(y_{t_j}^{\text{CPI}}I_v'\right) &= C_{\text{CPI}}\mathrm{E}(f_{t_j} - f_{t_j|\Omega_{v-1}}) \\ &\quad \times (f_{t_i} - f_{t_i|\Omega_{v-1}})'C_{\text{RMP}}' \quad \text{and} \\ \mathrm{E}\left(I_vI_v'\right) &= C_{\text{RMP}}\mathrm{E}(f_{t_i} - f_{t_i|\Omega_{v-1}}) \\ &\quad \times (f_{t_i} - f_{t_i|\Omega_{v-1}})'C_{\text{RMP}}' + \Sigma_{\text{RMP}},\end{aligned}$$

where Σ_{RMP} is a diagonal matrix with elements of Σ which correspond to the RMP data. The expectations $\mathrm{E}(f_{t_j} - f_{t_j|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'$ and $\mathrm{E}(f_{t_i} - f_{t_i|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'$ can be obtained from the Kalman filter. Consequently, we can find a vector B such that the following holds:

$$\underbrace{y_{t_j|\Omega_v}^{\text{CPI}}}_{\text{new forecast}} = \underbrace{y_{t_j|\Omega_{v-1}}^{\text{CPI}}}_{\text{old forecast}} + B\Big(\underbrace{y_{t_i}^{\text{RMP}} - y_{t_i|\Omega_{v-1}}^{\text{RMP}}}_{\text{news}}\Big). \tag{9}$$

This enables us to trace the sources of forecast revisions. More precisely, in the case of a simultaneous release of several (groups of) variables, it is possible to decompose the resulting forecast revision into contributions from the *news* in individual (groups of) series.[5] In addition, we can

---

[4] In what follows, we do not take changes in the parameter estimates into account. The influence of this factor needs to be analyzed separately.

[5] If the release concerns only one group or one series, the contribution of its *news* is simply equal to the change in the forecast.

produce statements such as "after the release of Raw Material Prices, the forecast of CPI inflation went up because the indicators turned out to be (on average) higher than expected".[6]

5.2. Do financial variables matter?

Eq. (9) shows that the unexpected part of the release, i.e. the part which was not predicted by the model, is the only part of a new release that revises the forecast of the target variable. In this section we use this finding to evaluate the extent to which financial variables matter in producing accurate backcasts, now-casts and forecasts of the total CPI inflation. The application refers to the financial variables included in our analysis, i.e., the trade-weighted US dollar index (major currencies), the *S&P* 500 index, the ten-year Treasury constant maturity rate, and the three-month Treasury-bill rate.

As a first step, we establish that the inclusion of sequential new information improves the forecast accuracy. In order to show this, we perform the following exercise: we evaluate the performance of the best model ex-post (i.e. the model with one factor and 6 lags, using all information available) at different points in time (i.e. on the 1st, the 7th, the 15th, the 21st and the 28th of each month), meaning that we use all of the information available on the specified days of the month. Following that, we average the squared errors of each day across months. Then, we evaluate the 1-month-ahead forecast and the now-cast for each of the specified days in the month. We also evaluate the backcast for the 1st and 7th day in the month (the release of the CPI is usually on the 15th of the month following the reference period, and therefore no further evaluation is needed). This will provide an indication of how these averages decline (i.e., how the forecast accuracy is improved) as new information becomes available.

Based on this, in order to show the value added by financial variables, we repeat the procedure with two variations, thus producing two counterfactual exercises. First, we neutralize the effect of financial variables. This can be understood as setting the *news* related to financial variables equal to zero in Eq. (9). In practice, this can be achieved by replacing the financial data released between two consecutive periods (e.g. the entire financial variable releases between the 1st and the 7th of each month) with the forecast produced by the model for those same days. According to Eq. (9), the impact of the financial release on the revision of total CPI inflation forecast will be zero, and any revisions will be due to the other variables included in our data set. In the second counterfactual exercise, we attempt the opposite, i.e., we allow the information from the financial variables to be the unique information that arrives in each time interval. In practice, we replace the releases of RMP and WRGDP between two consecutive evaluating dates with the forecast produced by the model. According to Eq. (9), the overall impact of all other releases on the revision of the total CPI inflation forecast will be zero, and the revisions will be due solely to the financial variables included in our data set. Finally, we compare the RMSFEs obtained from the first step with the ones produced by the two counterfactual exercises.

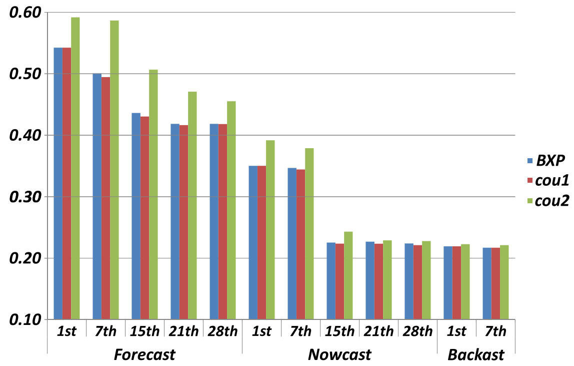

Fig. 9. Counterfactual exercise. *Notes*: This figure shows the evolution RMSFEs computed every week, from the month preceding the reference period, to the second week of the month following the reference period. *BXP* indicates the best model ex-post, i.e., the model that uses all of the available data with the parametrization of one factor and six lags. *cou1* are the RMSFEs produced assuming that all of the financial data released between two consecutive dates are equal to the forecast produced on the first date. *cou2* are the RMSFEs produced assuming that all of the RMP and WRGDP data released between two consecutive dates are equal to the forecast produced on the first date.

Fig. 9 shows the results of these exercises. First of all, it is important to notice that the RMSFEs decrease monotonically when new information arrives, in all three exercises. For example, the forecast produced on the 1st day of each month by the best model ex-post (*BXP*) has a RMSFE equal to 0.54, and this decreases monotonically to 0.22 on the 7th day of the backcast. This suggests that the data released each week contain important information which improves the forecast accuracy of the proposed model. Not surprisingly, the most relevant improvement in the forecast accuracy occurs between the 7th and the 15th of the current month (now-cast), when the CPI data for the previous month are released.

However, this figure also suggests that financial variables are of little, if any, value in forecasting the total CPI inflation. Indeed, looking at the RMSFEs of the first counterfactual exercise (*cou1*), it appears that they are similar to the RMSFEs produced by *BXP*. On the other hand, looking at the RMSFEs of the second counterfactual exercise (*cou2*), it becomes obvious that the forecast performance under *cou2* deteriorates sharply. This is due to the fact that, while financial variables are timelier than RMP and WRGDP, their co-movement with the total CPI inflation is not equally strong. Figs. 4–7 provide corroborating evidence on this front. Overall, the results of the counterfactual exercises suggest that, as long as the data set includes the RMP and WRGDP variables, the addition of financial variables does not appear to improve the forecasting accuracy of our model significantly.

6. Summary

This paper proposes an econometric framework that exploits weekly and daily data in order to forecast US total

[6] This holds of course for indicators with positive entries in *B*.

Table 4
Data.

Name	Source
Monthly	
CPI—All items	Bureau of Labor Statistics
CPI—Energy	Bureau of Labor Statistics
CPI—Food and Beverages	Bureau of Labor Statistics
CPI—Housing	Bureau of Labor Statistics
CPI—Other goods and services	Bureau of Labor Statistics
CPI—Transportation	Bureau of Labor Statistics
ISM Manufacturing: Prices Index	Institute for Supply Management
ISM Non-Manufacturing: Prices Index	Institute for Supply Management
Weekly	
Diesel Sales Price	Energy Information Administration
Midgrade All Formulations Gas Price	Energy Information Administration
Premium All Formulations Gas Price	Energy Information Administration
Regular All Formulations Gas Price	Energy Information Administration
Daily	
Food and tropical beverages	Org. for Economic Cooperation and Development
Cereals	Org. for Economic Cooperation and Development
Oilseeds and oil	Org. for Economic Cooperation and Development
Beverages, sugar and tobacco	Org. for Economic Cooperation and Development
Industrial raw materials	Org. for Economic Cooperation and Development
Agricultural raw materials	Org. for Economic Cooperation and Development
Spinning material	Org. for Economic Cooperation and Development
Non-ferrous metals	Org. for Economic Cooperation and Development
Iron ore, scrap	Org. for Economic Cooperation and Development
Coal	Org. for Economic Cooperation and Development
Crude oil	Org. for Economic Cooperation and Development
10-Year Treasury Constant Maturity Rate	Board of Governors of the Federal Reserve System
3-Month Treasury Bill: Secondary Market Rate	Board of Governors of the Federal Reserve System
S&P 500 Index	Standard and Poor's
Trade Weighted US Dollar Index: Major Currencies	Board of Governors of the Federal Reserve System

CPI inflation. The paper focuses on four groups of data, including sampling frequencies which are higher than monthly. The first group consists of the World Market Price of Raw Materials at a daily frequency. The second includes weekly surveys of fuel prices at the pump, the Weekly Retail Gasoline and Diesel Prices for the US. The third group includes some of the most important financial variables, namely the trade weighted US dollar index, the *S&P* 500 index, the ten-year Treasury constant maturity rate and the three-month Treasury-bill rate. Finally, the last group contains the ISM manufacturing and non-manufacturing prices indexes. The paper focuses on these data for two reasons: first, they co-move with inflation, and second, they become available in a more timely fashion than the inflation data. In addition to these four groups of data, the target forecast variables are also included, i.e., US total CPI inflation and its main subcomponents.

We model these data as a dynamic factor model estimated at daily frequency. Our estimation methodology is the one proposed by Bańbura and Modugno (2010), which allows us to estimate factor models on data sets with arbitrary patterns of missing data, e.g., with series sampled at different frequencies or with varying publication lags. Moreover, it allows us to introduce restrictions on the coefficients, such as the time-varying coefficients that this paper uses, in order to aggregate the daily factors. This framework offers unique estimation advantages over the previous literature. For example, it avoids the need to average high frequency data in order to obtain monthly frequency indicators, as did Lenza and Warmedinger (2011). This allows us to exploit the co-movement in our data set fully, without losing information. Moreover, in contrast to MIDAS models, e.g. Monteforte and Moretti (2010), the proposed framework does not require high frequency data to be imposed as predetermined regressors, and therefore it allows model-based *news* to be disentangled from each release and their impact on forecast revisions to be assessed.

The results suggest that the chosen weekly and daily data are important for improving the forecast accuracy for the US total CPI inflation, especially at the shortest horizon, i.e. now-casting. The model performs considerably better than either a model that uses only monthly information or standard benchmark models such as the random walk or an IMA(1, 1). This is due in particular to the improved forecast accuracy of the transport and energy components, which are the most volatile among the CPI growth rates. Moreover, the paper illustrates the use of the model in identifying *news* effects (i.e., the revisions of the forecast of the target variable that arise from new data releases), which enhances our understanding of which variables are crucial to the production of an accurate now-cast/forecast, thus further emphasizing the potential of this framework as an important tool for policy analysis. In particular, we show that the inclusion of high frequency data on energy and raw material prices in our data set contributes considerably to the gradual improvement in model performance. As long as these data sources are included in our data set, the inclusion of financial variables does not make any considerable improvement to the now-casting accuracy.

Acknowledgments

The author would like to thank Marta Banbura, Domenico Giannone, Michele Lenza, Kleopatra Nikolaou and Lucrezia Reichlin for their comments, as well as the seminar participants at the Board of Governors of the Federal Reserve System, CFE-ERCIM 2010, EEA-ESEM 2012, European Central Bank, EuroStat Colloquium 2010, ISF 2011, IWH/CIREQ macroeconometric workshop 2011 and the Norges Bank, Modeling and forecasting oil prices workshop 2012. This work was supported by the "Action de recherche concertée" contract ARC-AUWB/2010-15/ULB-11.

Appendix. Data

See Table 4.

References

Angelini, E., Bańbura, M., & Rünstler, G. (2007). *Estimating and forecasting the Euro area monthly national accounts from a dynamic factor model.* Discussion paper. European Central Bank.

Bańbura, M., & Modugno, M. (2010). *Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data.* Working paper series: Vol. 1189. European Central Bank.

Bańbura, M., & Rünstler, G. (2007). *A look into the factor model black box. Publication lags and the role of hard and soft data in forecasting GDP.* Working paper series: Vol. 751. European Central Bank.

Banerjee, A., & Marcellino, M. (2006). Are there any reliable leading indicators for US inflation and GDP growth? *International Journal of Forecasting*, *22*(1), 137–151.

Cecchetti, S. G., Chu, R. S., & Steindel, C. (2000). The unreliability of inflation indicators. *Current Issues in Economics and Finance*, *6*(4).

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

Doz, C., Giannone, D., & Reichlin, L. (2006). *A quasi maximum likelihood approach for large approximate dynamic factor models.* Working paper series: Vol. 674. European Central Bank.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). *The MIDAS Touch: Mixed Data Sampling Regression Models.* University of California at Los Angeles, Anderson Graduate School of Management qt9mf223rs, Anderson Graduate School of Management, UCLA.

Giannone, D., Henry, J., Lalik, M., & Modugno, M. (2012). An area-wide real-time database for the Euro area. *The Review of Economics and Statistics*, *94*(4), 1000–1013.

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

Lenza, M., & Warmedinger, T. (2011). A factor model for Euro-area short-term inflation analysis. *Journal of Economics and Statistics* (Jahrbuecher fuer Nationaloekonomie und Statistik), *231*(1), 50–62.

Mariano, R., & Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, *18*, 427–443.

Monteforte, L., & Moretti, G. (2010). *Real time forecasts of inflation: the role of financial variables.* Working paper series: Vol. 81. LUISS.

Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, *44*, 293–335.

Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, *41*(3), 788–829.

Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*(S1), 3–33.

Watson, M. W., & Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, *23*, 385–400.

Michele Modugno is a Post-Doc researcher at Université libre de Bruxelles. His research interests include time series analysis and applied macroeconomics. He has published in Journal of Applied Econometrics and the Review of Economics and Statistics.