New Eurocoin: A Tutorial Note

The following note provides a simple presentation of ideas and methods underlying the New Eurocoin indicator. For a detailed treatment see Altissimo, F., Cristadoro, R., Forni, M., Lippi, M. and G. Veronese (2008) “New Eurocoin: Tracking Economic Growth in Real Time”: see `http://www.cepr.org/pubs/new-dps/dplist.asp?dpno=5633`.

1. Measuring aggregate economic activity in the Euro area.

The most important official indicator of aggregate economic activity is the GDP. Unlike IP (Industrial Production), GDP includes Services, Agriculture and the Public Sector. Unlike the Surveys, GDP does not contain any subjective judgement. However, unlike IP and Surveys, GDP is released only quarterly and with delay. Lastly, IP, Surveys and GDP, all contain short-run oscillations, so that, for example, the beginning of a medium-run upswing cannot be distinguished from an improvement lasting for just two or three months.

The Eurocoin indicator combines the positive aspects of GDP, IP and Surveys:

a. It is *comprehensive,* in that, like GDP, it takes into account Production, Services and the Public Sector.

b. It is released monthly and is *timely:* around the 20-th of each month we are able to produce a reliable estimate for the previous month.

Moreover, Eurocoin is free from short-run fluctuations.

2. The medium- to long-run component of a macroeconomic variable.

2.1 Although there may be disagreement about which macroeconomic series is best suited to represent economic activity, it is a well established fact that to judge the current economic outlook and analyse the business cycle behaviour the chosen series must be firstly detrended. If z_t is the selected series, alternative detrending transformations are:

(i) Taking the residuals of a regression over time:

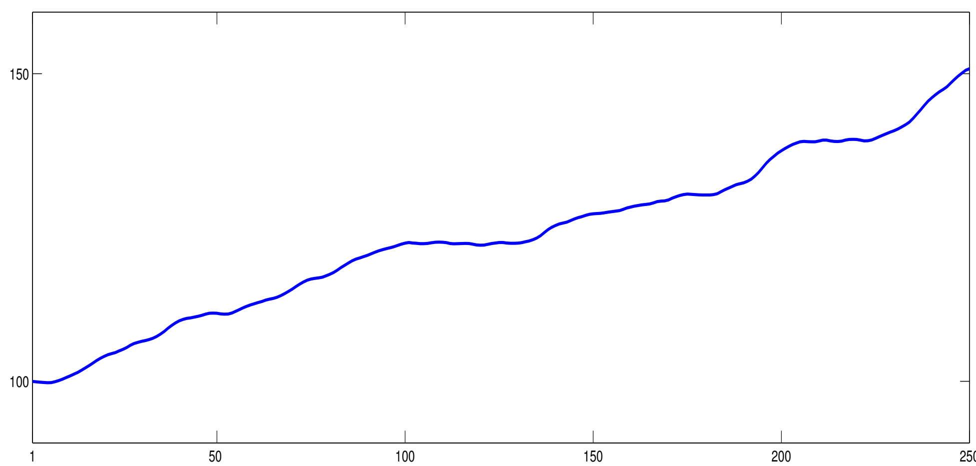
$$\log z_t = a + bt + u_t.$$

(ii) Taking the difference:

$$\log z_t - \log z_{t-1} = b + v_t.$$

Whether the first or the second transformation should be used is a long story. It will be sufficient here to recall that the second has definitely won the race, so that it is by now common practice to describe macroeconomic time series as non stationary (trending) series such that their rate of growth (which is approximately the log difference in the second equation above) is stationary.

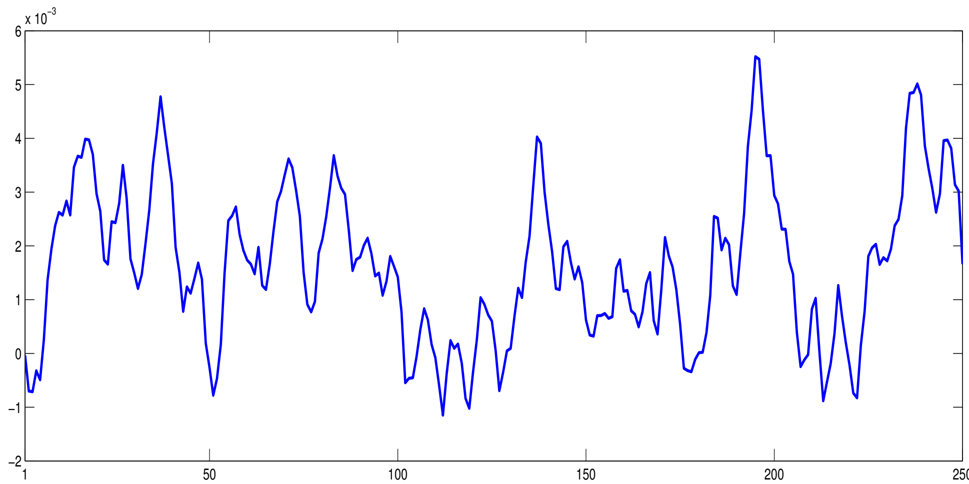
Figure 1: An artificial macroeconomic time series



The artificial series plotted in Figure 1 exhibits some of the typical features of macroeconomic time series. It is trending, though with a varying rate of growth. This, i.e. $y_t = \log z_t - \log z_{t-1}$, is plotted in Figure 2. The series y_t is clearly stationary, with an average of 0.16%. If the data are interpreted as monthly, then the log difference y_t is approximately the monthly rate of change of z_t, the figure 0.16% implying an yearly average rate of change of around 2%.

The plot in Figure 2 provides a good illustration of the idea that every stationary time series can be thought of as the sum of waves of different periods (this is formalized in the Spectral Representation Theorem). Long waves, with period between three and five years (36 to 60 months) are clearly discernible, as well as waves whose period is shorter than 12

Figure 2: The rate of growth of z_t: $y_t = x_t - x_{t-1}$



months. For example, a complete 59-monthts cycle, from peak to peak, occurs between the months 137 and 195. However, an economist or a policy maker, observing the rate of growth *during that period* and trying to decide whether the economy is still slumping or recovering, would have to change his/her assessment quite often. As a matter of fact, we see a considerable number of peaks and troughs corresponding to short oscillations. But neither the economist nor the policy maker is interested in these temporary phenomena. Our aim in the construction of Eurocoin is the removal of the waves whose period is one year or shorter from the rate of growth of z_t, without distorting the residual medium-long-run component. To illustrate our construction let us discuss a popular measure of medium- long-run growth, that is the year-on-year rate of change $\log z_t - \log z_{t-12}$. Setting $x_t = \log z_t$, observe that

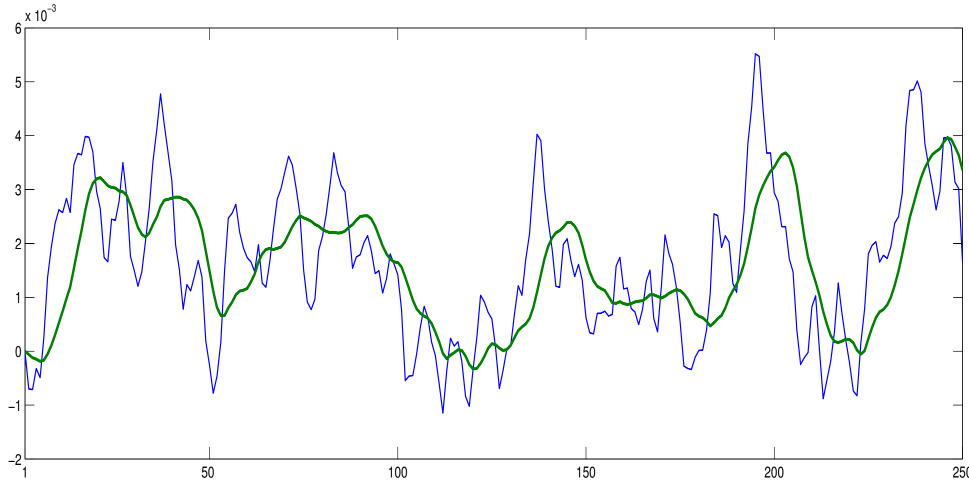
$$\frac{1}{12}(x_t - x_{t-12}) = \frac{1}{12}[(x_t - x_{t-1}) + (x_{t-1} - x_{t-2}) + \cdots + (x_{t-11} - x_{t-12})] \qquad (1)$$

(we divide by 12 to make its size comparable with the month-on-month figure). The year-on-year change at t and at $t-1$ are twelve-terms *moving averages* with eleven terms in common, thus it is very smooth as compared to the month-on-month change.

This smoothing effect is shown in Figure 3, in which y_t and $\frac{1}{12}(x_t - x_{t-12})$ are plotted together. We see that removal of the short-run oscillations makes it possible to clearly observe shorter and longer cycles in the series, and the corresponding turning points.

However, the year-on-year change has a serious drawback. As it is apparent in Figure 3, using the year-on-year difference causes a phase distortion, i.e. the peaks and troughs

Figure 3: Plot of $y_t = x_t - x_{t-1}$ (blue line) and $\frac{1}{12}(x_t - x_{t-12})$ (green line)



of the waves clearly occur with a delay. The reason can be easily grasped examining (1): The moving average on the right hand side only contains lags of $x_t - x_{t-1}$. In order to keep the phase of the transformed series, the moving average should be centered and symmetric. The moving average

$$x_{t+6} - x_{t-6} = \frac{1}{12}[(x_{t+6} - x_{t+5}) + \cdots + (x_t - x_{t-1}) + \cdots + (x_{t-5} - x_{t-6})]$$

would be almost good but not completely, indeed it has 6 leading, one central and 5 lagging terms, whereas the average $\frac{1}{2}\left[\frac{1}{12}(x_{t+6} - x_{t-6}) + \frac{1}{12}(x_{t+5} - x_{t-7})\right]$, that is

$$w_t = \frac{1}{24}(x_{t-6}-x_{t-7})+\frac{1}{12}[(x_{t-5}-x_{t-6})+\cdots+(x_t-x_{t-1})+\cdots+(x_{t+5}-x_{t+4})]+\frac{1}{24}(x_{t+6}-x_{t+5}) \quad (2)$$

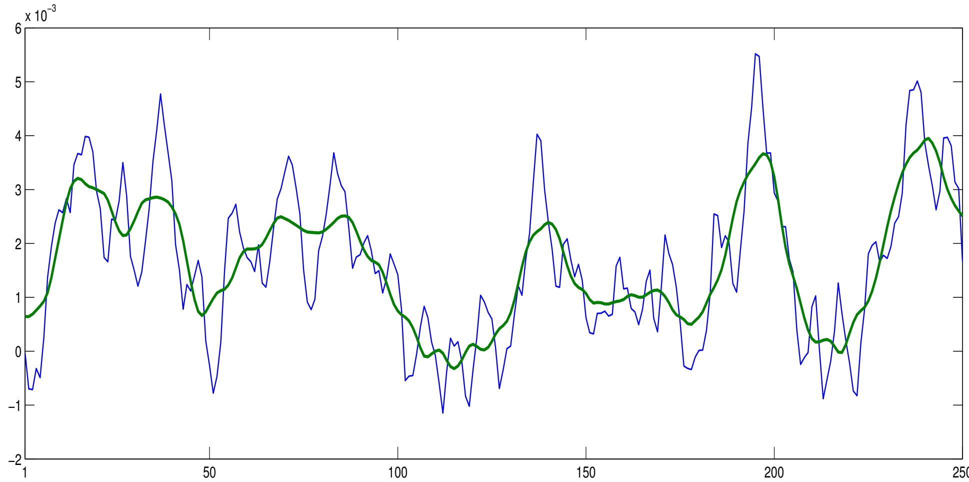
is both central and symmetric.

The result of using (2) instead of (1) is shown in Figure 4. As compared to Figure 3, the green line is shifted backward (and is slightly smoother), so that the phase distortion disappears.

However, if T indicates the end of the sample, the values of $x_t - x_{t-1}$ with $t > T$ are not observed and must be replaced by forecasts. As new information arrives, that is at times $T+1,\ T+2,\ \ldots,\ T+6$, the average (2), computed for T (but also for $T-1,\ \ldots,\ T-5$), will have to be revised.

In conclusion, using two-sided moving averages like (2) to smooth out short-period oscillations runs into difficulties at the end of the sample, the reason being that some of

Figure 4: As in Figure 3, with w_t instead of $\frac{1}{12}(x_t - x_{t-12})$



the values that are necessary must be replaced by predictions. This problem is discussed in detail in the next section. On the other hand, solving this end-of-sample problem is crucial for the construction of a smooth, *timely* indicator of economic activity.

2.2 The moving average w_t is a fairly good solution to the problem of removing oscillations of period shorter then one year while keeping the remaining oscillations in phase. However, as time-series theory shows, the weights used in (2) are not exactly what is needed to avoid any undesired alterations in the time series. The "correct" moving average is known as *band pass filter*. It is infinite and uses weights β_j that tend to zero as j tends to infinity. Denoting by c_t the band-passed series,

$$c_t = \cdots + \beta_k y_{t-k} + \cdots + \beta_0 y_t + \cdots + \beta_k y_{t+k} + \cdots \tag{3}$$

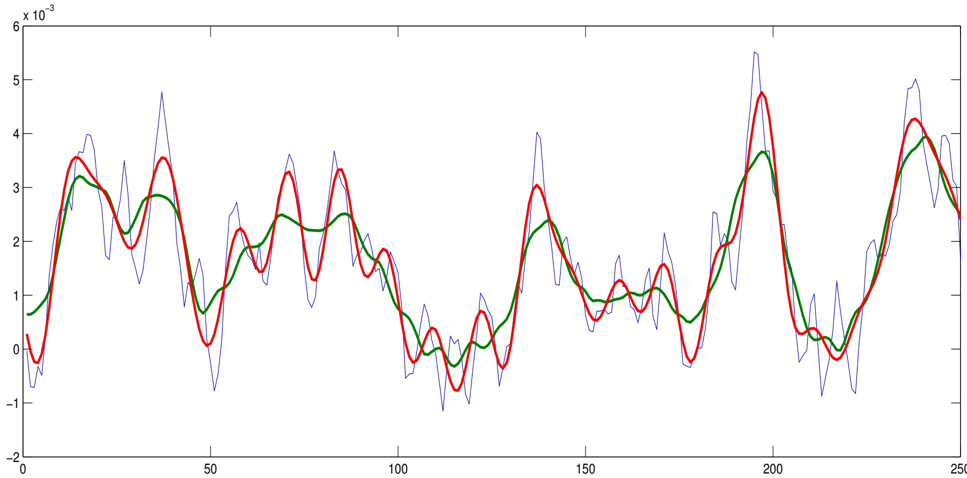
where

$$\beta_k = \begin{cases} \dfrac{\sin(k\pi/6)}{k\pi} & \text{for} \quad k \neq 0 \\ 1/6 \quad \text{for} \quad k = 0. \end{cases}$$

The application of (3) to our artificial series is shown in Figure 5 (red line) together with y_t and w_t (blue and green respectively). We see that c_t is very smooth. Moreover, comparison between c_t and w_t shows that w_t removes "too many" frequencies, i.e. that some of the waves longer than one year are erased or substantially reduced.

However, though being extremely precise as regards the waves to be removed, the band pass filter is another centered and symmetric moving average (in order to keep the phase),

Figure 5: Plot of y_t (blue line), w_t (green line) and c_t (red line)



so that we are left with the same end-of-sample problem we have mentioned in the conclusion of the previous section. Precisely, computing the moving average (3) requires that the values of y_t that are missing, corresponding to $t > T$ and $t < 1$, be replaced by forecasts (backcasts for $t < 1$). The series c_t plotted in Figure 5 is based on the most naïve forecast of the missing values, namely $y_{T+k} = y_{1-k} = \bar{y}$, where $\bar{y}$ denotes the sample average of y_t. It is easily seen that as soon as new information arrives, at $T+1$, $T+2$, ..., all the values of the series c_t change. But if we consider, say, $T-100$, the change is insignificant, whereas serious revisions may occur at the end of the sample.

As an illustration, consider again the artificial series y_t and c_t, between 1 and 250. In Figure 6:

a. The red line represents the band passed series computed using the data up to 250, call it $c_t^{(1)}$ (this is the last part of the red line in Figure 5).

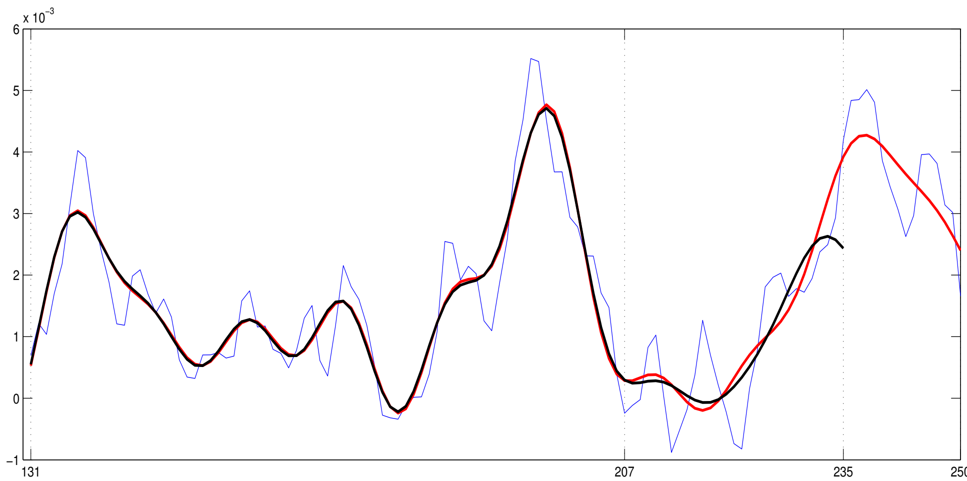
b. The black line represents c_t computed when $T = 235$, i.e. with the data between 236 and 250 missing and therefore replaced by the average of y_t, call it $c_t^{(2)}$.

Note that $c_t^{(1)}$ is computed using the average of y_t for $t > 250$, thus its values between, say, 240 and 250 must be taken with care. However for $t \leq 235$, $c_t^{(1)}$ does not change much as new data arrive, for $t > 250$, so that $c_t^{(1)}$ can be considered as a reliable estimate of c_t for $t \leq 235$.

We see in Figure 6 that a small difference between $c_t^{(2)}$ and c_t, i.e. between the black and the red line, emerges after $t = 207$. However, at the end of the sample the difference becomes very serious, with the black line signaling a turning point that is definitely "false", as both the red line and the blue line, i.e. y_t, clearly show.

In conclusion, for values of t within the sample, say for 12 periods away from T, the band-passed series, obtained by setting the missing values equal to the average of y_t, is a reliable estimate of c_t, with insignificant revisions as new data arrive. On the contrary, as the above example shows, at the end of the sample the estimate of c_t may be seriously misleading.

Figure 6: y_t between 131 and 250 (blue line). c_t between $t = 131$ and $t = 250$ (red line). c_t computed when only data between $t = 1$ and $t = 235$ are available (black line).



3. New Eurocoin. A sketch of the method.

New Eurocoin is a solution to the problem just outlined, that is obtaining a reliable estimate of c_t at the end of the sample, i.e. a good assessment of the medium- long-run component of y_t in real time.

As described above, to compute c_t we have used the average of y_t as predictor for out-of-sample values. Obviously we can do better by resorting to more sophisticated techniques:

(A) Univariate ARMA models. Fitting an ARMA to y_t we obtain predictions of y_t for $t > T$, which improve upon the sample average as soon as y_t has non trivial autocorrelation (this is apparently the case with our simulated series).

(B) Small-size multivariate ARMA models. We can improve upon univariate ARMA models by modeling y_t together with other macroeconomic series.

(C) Large factor models. When a large dataset is available, say hundreds of macroeconomic time series, factor techniques can be applied to improve the prediction. Factors are linear combinations of the variables in the dataset. A small number of them can be used as a very good summary of the relevant information. The following is an elementary example. Suppose that the variables of the dataset follow the model

$$x_{it} = au_t + bu_{t-1} + \xi_{it}, \tag{4}$$

where
$i = 1, 2, \ldots, n$,
u_t is a white noise scalar (a process with no autocorrelation),
ξ_{it} is also a white noise, the difference with respect to u_t being that u_t is common to all variables while ξ_{it} is specific to variable i,
u_t and ξ_{it} are uncorrelated, for all i,
ξ_{it} and ξ_{jt} are uncorrelated for all i and j; assume for simplicity that the variance of ξ_{it} is independent of i, and call it σ_ξ^2.
The "macrovariables" x_{it} are driven by a *common component*, that is $au_t + bu_{t-1}$, plus a specific component, called *idiosyncratic* in the literature. This extremely stylized model provides quite a clear idea as to how the factor structure can be employed for forecasting. Let us now take the average of the x's:

$$\bar{x}_t = \frac{1}{n}\sum_{i=1}^{n} x_{it} = au_t + bu_{t-1} + \frac{1}{n}\sum_{i=1}^{n} \xi_{it},$$

and its variance

$$\mathrm{var}(\bar{x}_t) = (a^2 + b^2)\sigma_u^2 + \frac{1}{n^2}\sum_{i=1}^{n} \sigma_\xi^2 = (a^2 + b^2)\sigma_u^2 + \frac{1}{n}\sigma_\xi^2$$

(this is an elementary calculation, based on the fact that the variance of uncorrelated variables is the sum of their variances).
Therefore, when n is large $\bar{x}_t$ is almost equal to the common component $au_t + bu_{t-1}$. Assuming that y_t is the first variable in the dataset, that is $y_t = x_{1t}$, we can predict y_t by separately predicting $\bar{x}_t$, which is a moving average of order one, and the idiosyncratic component ξ_{1t}. This provides a clear advantage over both a univariate ARMA and small multivariate ARMA models. In both cases, the moving average $au_t + bu_{t-1}$ gets mixed with the idiosyncratic components.
(D) The construction of Eurocoin is based on a factor model like that exemplified above in (4). However, we do not make use of explicit predictions of y_t, $t > T$. Rather,

Eurocoin is obtained as a projection of the (bandpassed) GDP on factors obtained through a generalization of the concept of principal components. An illustration of our idea is given below using, again, a stylized model.

As above, assume that the variables in the dataset are x_{it} and that $y_t = x_{1t}$. Moreover, let

$$x_{it} = u_{t-k_i} + \xi_{it} \tag{5}$$

for $i = 2, \dots, n$, so leaving aside for the moment y_t, and let k_i take the values 0, 1 or 2. Thus we have three subgroups, the variables that are *leading*, those loading u_t (corresponding to $k_i = 0$), the variables that are *lagging*, those loading u_{t-2} ($k_i = 2$), and the *central* variables, those loading u_{t-1} ($k_i = 1$).

Lastly, assume that y_t is central: $y_t = u_{t-1} + \xi_{it}$. For the sake of simplicity let us use a truncation of the band-pass filter (3):

$$\kappa_t = \beta_1 y_{t-1} + \beta_0 y_t + \beta_1 y_{t+1}. \tag{6}$$

We have:

$$\kappa_t = [\beta_1 u_{t-2} + \beta_0 u_{t-1} + \beta_1 u_t] + [\beta_1 \xi_{1,t-1} + \beta_0 \xi_{1t} + \beta_1 \xi_{1,t+1}] = U_t + V_t. \tag{7}$$

Within the sample, for $t < T$, the moving average κ_t can be computed using (6). But, at $t = T$, (6) is no longer feasible. However, the three terms adding up to U_t can be recovered using the *current* values of the variables in the dataset. In particular, u_t can be obtained as an average of current values of the leading variables. If we can assume that the idiosyncratic component is a fairly small fraction of the GDP, then recovering U_t at the end of the sample represents a good approximation of κ_t.

We cannot go into further details of our construction here. The basic idea, illustrated in the above example, is that variables belonging to the dataset that are leading with respect to the GDP act as proxies for future values of the GDP, that are necessary to obtain c_t but are missing at the end of the sample. Our method can be summarized as follows:

(I) We construct m linear combinations of *current values* of the variables x_{it} that automatically select lagging, central and *leading* variables in order to maximize smoothness. The number m is determined by a compromise between smoothness and goodness of fit. Our linear combinations, a generalization of ordinary principal components, are referred to as *smooth factors* and denoted by $F_{1t},\ F_{2t},\ \dots,\ F_{mt}$.

(II) Then we regress c_t over the factors:

$$c_t = A_1 F_{1t} + A_2 F_{2t} + \cdots + A_m F_{mt} + R_t \tag{8}$$

(the mean has been subtracted from all variables, so that no constant appears in the regression). The covariances necessary to estimate the coefficients A_j can be computed using values of c_t that are far from the beginning and the end of the sample, where (3) provides a good estimate.
(III) Lastly, the value of c_t at the end of the sample is estimated using (8):

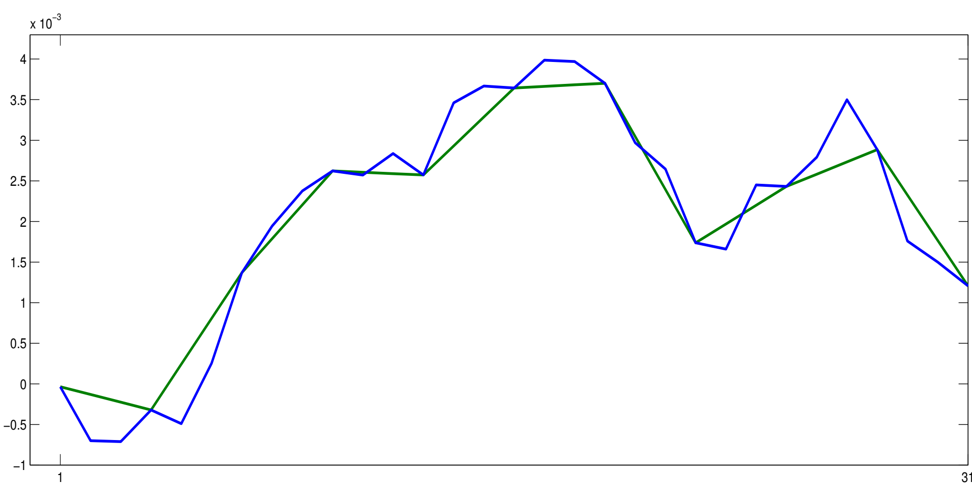
$$\hat{c}_T = A_1 F_{1T} + A_2 F_{2T} + \cdots + A_m F_{mT}. \tag{9}$$

This is the Eurocoin equation. It provides an estimate of c_t at the end of the sample which is not based on predicted values of y_t but on *current values* of the variables x_{it}, and is therefore feasible at T.
(IV) Note that, unlike the estimates of c_t that are based on predictions of future missing observations, the performance of

$$A_1 F_{1t} + A_2 F_{2t} + \cdots + A_m F_{mt},$$

as an estimate of c_t, does not depend on whether t is far, close or equal to T. However, the advantage of New Eurocoin with respect to other estimates of c_t shows up at T, that is in *real time*, whereas for t far from T estimates based on predicted missing values are clearly superior.

Figure 7: Plot of y_t (blue line) and $\tilde{y}_t$ (green line) for $t = 1, \ldots, 31$



4. Empirical implementation of New Eurocoin.

4.1 For the sake of clarity the previous section gives a stylized presentation of New Eurocoin. In particular, we have assumed that y_t is observed monthly. In practice however the GDP is released only quarterly. As a consequence, obtaining the monthly series c_t, which is on the left hand side of the projection (8), requires some further work.

A convenient way to deal with this problem is to think of the variable z_t as aggregating the GDP flow over the months t, $t-1$, $t-2$, with two observations missing every quarter. Thus, for example, if t is January 1987, then z_t is the aggregate over November 1986, December 1986, January 1987, and is a missing observation, whereas z_{t-1}, corresponding to December 1986, is the ordinary observed figure for the fourth quarter of 1986.

Moreover, y_t is not defined as $\log z_t - \log z_{t-1}$, which would be never observable, but as the monthly quarter-on-quarter rate of growth

$$y_t = \log z_t - \log z_{t-3},$$

which has, just like y_t, two missing observations every quarter. Techniques to obtain monthly observations for a series like y_t range from linear interpolation, which simply draws the straight segment joining subsequent observed figures, to sophisticated regressions of the series with missing observations over monthly reference indicators (see Chow and Lin, 1971, see the references in Altissimo et al., 2008).

However, in our case things are simplified by the fact that our focus is not on y_t but on c_t. The latter is obtained by removing short-run oscillations from y_t, so that the particular method employed to interpolate values within quarters matters very little. A simple experiment will illustrate this point.

Start with the artificial series y_t, remove two observations for every quarter, so that we are left with $y_1,\ y_4,\ y_7,\ \ldots$, then replace the missing observations by the linear interpolation values, call $\check{y}_t$ the resulting monthly series.

Then construct the band-pass filtered series using $\check{y}_t$ in place of y_t, call it $\check{c}_t$. The ratio $\mathrm{var}(c_t - \check{c}_t)/\mathrm{var}(c_t)$ is equal to 1.4%, thus using the interpolated series $\check{y}_t$ in place of the "real" series y_t has a negligible effect after band passing.

4.2 As mentioned at the beginning of the present note, a detailed account of the performance of New Eurocoin is given in Altissimo et al. (2008). The indicator published in this website is based on the work done in that paper but improves on it in at least two ways: (i) the data set on which the indicator published here is based has been upgraded substituting series that were updated with long delays with more timely, equivalent ones and enlarged to comprehend also daily variables (ii) the indicator for a given month t is now available before the end of that month (and not by the 20th of the following month as in the paper. Let us summarize here the key features of the dataset employed and some performance indicators.

(I) The dataset includes about 150 series from Thomson Financial Datastream, referring to the euro area as well as its major economies, and starts in June 1987. The dataset includes groups of variables that are, according to current practice in conjunctural analysis, leading, lagging and coincident with respect to the GDP. In particular, the presence of leading variables, which contain information about future values of the GDP, is crucial to obtain a good estimate of c_t at the end of the sample. The database is organized into homogeneous blocks, i.e. industrial Production Indexes (41 series), Prices (24), Money Aggregates (8), Interest Rates (11), Financial Variables (6), Demand Indicators (14), Surveys (25), Trade Variables (9) Labour Market Series (7) and daily series (stock markets).

(II) The variables in the dataset are available with different delays. Around the 20-th of month T, when the indicator for month T is produced, Surveys, daily sereis and Financial Variables are usually available up to time T, thus with no delay, Car Registrations and Industrial Orders up to $T-1$ and Industrial Production indexes up to $T-2$ or $T-3$. This end-of-sample unbalance problem can be solved by forward realignment, that is by simply shifting forward the variables that are available with delay. As an alternative, the variables that are not available might be predicted, using ARMA models or the EM algorithm. We are currently using all these methods and monitoring the results.

(III) Regarding the performance, we show in the paper that the R^2 of regression (8) is 0.79, and that the ratio of correct predictions of the sign of $c_t - c_{t-1}$ by the sign of $\hat{c}_t - \hat{c}_{t-1}$ is 88%. Moreover, a real-time exercise shows that $\hat{c}_t$ does provide a good indicator of the turning points in c_t. Analogous results are obtained with the new dataset and the estimation at T on T.