Ex Post and Ex Ante Analysis of Provisional Data *

Giampiero M. Gallo

Dip. di Statistica "G. Parenti"
Università di Firenze
Viale G.B. Morgagni, 59
I-50134 Firenze FI
tel. +39 55 423 7257
FAX +39 55 422 3560
and
Economics Department
European University Institute
gallog@stat.ds.unifi.it

Massimiliano Marcellino

IGIER
Università Bocconi
Via Salasco, 5
I-20136 Milano MI
tel. +39 2 5836 3327
FAX +39 2 5836 3302
and
Economics Department
European University Institute
massimiliano.marcellino@uni-bocconi.it

Revised: November 1998
JEL Classification: C53, E47, E51

*We would like to thank Neil Ericsson for kindly providing the data on M1 used in this paper. We benefitted from discussions with Glenn Rudebusch. While the customary disclaimer applies, we would like to thank Fabio Canova and two anonymous referees for their comments which helped streamline the presentation and better focus on the issues. Financial support from the Italian MURST and CNR is kindly acknowledged by Gallo.

Ex Post and Ex Ante Forecasting with Provisional Data

Ex Post and Ex Ante Forecasting with Provisional Data

Abstract

In this paper we suggest a framework to assess the degree of reliability of provisional estimates as forecasts of final data, and we reexamine the question of the most appropriate way in which available data should be used for ex ante forecasting in the presence of a data revision process. Various desirable properties for provisional data are suggested, as well as procedures for testing them, taking into account the possible nonstationarity of economic variables. For illustration, the methodology is applied to assess the quality of the US M1 data production process and to derive a conditional model whose performance in forecasting is then tested against other alternatives based on simple transformations of provisional data or of past final data.

Information about macroeconomic variables is collected and processed by agencies which release provisional figures and later revise them until they are considered "final", that is, not in need of further revisions. The process of convergence to finalized data may take a long time, although later ordinary revisions are usually of lesser importance. The impact that such provisional data have on economic activity is quite relevant: consider, for example, the effects that announcements for money supply, inflation or GNP have on the expectation climate and therefore on investment decisions and financial markets.

As the rational expectation literature has emphasized, the impact of an announcement is relevant only if it is unexpected, i.e., if it constitutes a surprise relative to an information set. Thus, from an empirical point of view, the correct evaluation of what a surprise is and of its impact hinges on a correct specification of the expectation formation process for the variable of interest, conditional on the information *currently* available. In fact, it is unrealistic to assume that final data are available without any delay, or that agents wait for their release before deciding which actions to take. Hence the need for a proper framework to evaluate the properties of provisional data as forecasts of final data.

In defining the surprises, a distinction has been made in the literature between unanticipated and unperceived movements of a macroeconomic variable. In reference to money supply, for example, *unanticipated* money growth is usually taken to be the difference between an extrapolation of past behavior of money growth and actual current money growth (final data), whereas *unperceived* money supply is the difference between preliminary and final values. Barro and Hercowitz (1980) find that if unperceived money growth is used in the model instead of unanticipated money growth, it loses all significant explanatory power for unemployment and output (cf. also Boschen and Grossman, 1982, for similar conclusions).

The fact that timely published data contain errors (which will be corrected at a later stage) should also be taken into account. For example, provisional data might signal a deviation in monetary policy even when such a deviation is not present and, as noted by Maravall and Pierce (1986), attempts at correcting such a deviation can insert noise into the system.

A further reason for studying the information contained in provisional data relative to final data is to evaluate the "rationality" of the data production pro-

cess. The possibility of increasing the accuracy of provisional data by using already available information would make it convenient for the agents to revise provisional data themselves, instead of relying on officially published data.

The consequences of the presence of provisional data have long been investigated in the literature: previous studies focus on descriptive assessments of the quality of provisional data and their effects on estimation and forecasting with large-scale and time series models. Another stream of literature with which this study concords is concerned with real-time forecasting (Diebold and Rudebusch, 1991), which takes into explicit account the fact that at the time of performing a forecast the most recent data available are provisional.

In this paper we set the problem in more general terms, suggesting a procedure which addresses the two fundamental issues at hand, namely, the statistical properties of a given data production process and the possibility of improving upon published provisional data with an interest to the forecast of final data.

The main novelty of our procedure is that it considers explicitly the non-stationarity of most macroeconomic variables and the stationarity of revision errors (i.e. provisional and final data are cointegrated). By neglecting this aspect, one may misspecify the model used to assess the properties of provisional data and therefore obtain unreliable results.

The procedure can be easily adapted to study the relationship between all anticipating variables such as forward rates of exchange rates, futures rates, leading indicators, and so on, and their realized counterparts. The idea of using cointegration analysis in this context is not new; see, for example, Hakkio and Rush (1989), Patterson and Heravi (1991) and Hamilton and Perez-Quiros (1995). The original methodological aspects of this paper lie in having cast properties and procedures into a more formal framework.

The structure of the paper is as follows: in Section 1 we develop the econometric framework based on cointegration which will be used throughout the paper. In Section 2 we apply the methodology to monthly data for US M1. The procedure for forecasting final data from currently available provisional values is introduced in Section 3, and is then applied to the series at hand. Concluding remarks follow.

1 The Econometric Methodology

In what follows, we will simplify somewhat the complex reality of the various data production processes. Extensions of the analysis to actual cases can be notationally burdensome, but can easily be adapted from our framework. We will assume that preliminary figures, revisions and final data are published at regular intervals. This is in agreement with recent common practice by data production agencies. We will assume also that final data are the outcome of a process of successive data revisions.

Given a finite number of revisions, n, the sequence of data available through time for the value of a variable y_t at time t can be represented as:

$$_{t+1}p_t,\ _{t+2}r_t^1,\ _{t+3}r_t^2,\ \ldots,\ _{t+n}r_t^{n-1},\ _{t+n+1}f_t,$$

where we have indicated by $_{t+1}p_t$ the preliminary value for period t which becomes available in period $t+1$; $_{t+1+i}r_t^i$ is the i^{th} revisions for y_t which become available in period $t+1+i$, $i=1,\ldots,n \Leftrightarrow 1$ and $_{t+1+n}f_t$ is the final value available $n+1$ periods after t.

At each period, then, a number of preliminary, revised, and final data are announced for the series of interest. For example, taking time $t+1$ as a reference, the values

$$_{t+1}p_t,\ _{t+1}r_{t-1}^1,\ _{t+1}r_{t-2}^2,\ \ldots,\ _{t+1}r_{t-n}^{n-1},\ _{t+1}f_{t-n-1},$$

are published.

As the number of revisions increases, stylized facts suggest that it is unlikely that informative changes occur; hence, considering successive revisions is less relevant than concentrating just on first published data and first revisions. For this reason, and also to simplify the notation, we will assume throughout that $n=2$.

In order to characterize the nature of the relationship between provisional and final data from an *ex post* point of view, the relevant variables to be considered are

$$p_t \equiv_{t+1} p_t, \quad r_t \equiv_{t+2} r_t^1, \quad f_t \equiv_{t+3} f_t, \qquad t=1,\ldots,T \Leftrightarrow 2.$$

When these variables are integrated of order 1, $I(1)$, cointegration between provisional and final data is a necessary condition for the data to be of interest,

since large and systematic discrepancies could suggest either unreliability of data collection and processing, or an attempt at "fooling" the public.

Other properties to be examined relate to whether provisional data be considered Minimum Mean Squared Error Predictors of final data (or whether there exist some combination of current and past provisional and final data having this property) and to whether provisional data are unbiased forecasts of the final data. To do this, let us assume that a suitable statistical representation for $\{f_t, p_t\}_{t=0}^{\infty}$, is given by a $VAR(q)$

$$\mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{e}_t \tag{1}$$

where $\mathbf{y}_t = (f_t, p_t)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, $\mathbf{e}_t \sim i.i.d.N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \{\sigma_{ij}\}, i, j = 1, 2$, is positive definite and $\mathbf{A}(L) = \{a_{ij}(L)\} = (\mathbf{I} \Leftrightarrow \mathbf{A}_1 L \Leftrightarrow \mathbf{A}_2 L^2 \Leftrightarrow \ldots \Leftrightarrow \mathbf{A}_q L^q)$ is a matrix polynomial in the lag operator L. We will keep the relevant initial values fixed.

A common reparameterization of (1) yields the basis for cointegration testing,

$$\mathbf{B}(L)\Delta\mathbf{y}_t = \Leftrightarrow\mathbf{A}(1)\mathbf{y}_{t-1} + \boldsymbol{\mu} + \mathbf{e}_t \tag{2}$$

where $\Delta = (1 \Leftrightarrow L)$ is the first-difference operator, $\mathbf{B}(L) = (\mathbf{I} \Leftrightarrow \mathbf{B}_1 L \Leftrightarrow \mathbf{B}_2 L^2 \Leftrightarrow \ldots \Leftrightarrow \mathbf{B}_{q-1} L^{q-1})$ is a matrix polynomial of order $q \Leftrightarrow 1$, with $\mathbf{B}_i = \Leftrightarrow \sum_{j=i+1}^{q} \mathbf{A}_j$.

If f_t and p_t are cointegrated (Engle and Granger, 1987; Johansen, 1995), that is, if

$$\mathbf{A}(1) = \boldsymbol{\alpha}\boldsymbol{\beta}' = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} (1 \ \beta_1),$$

then we can write (2) as the restricted Vector Error Correction Model (VECM)

$$\mathbf{B}(L)\Delta\mathbf{y}_t = \Leftrightarrow\boldsymbol{\alpha} z_{t-1} + \mathbf{d} + \mathbf{e}_t. \tag{3}$$

where

$$z_t = \beta_0 + f_t + \beta_1 p_t. \tag{4}$$

and

$$\begin{aligned} \mathbf{d} &\equiv \boldsymbol{\beta}_{\perp}(\boldsymbol{\alpha}'\boldsymbol{\beta}_{\perp})^{-1}\boldsymbol{\alpha}'_{\perp}\boldsymbol{\mu}, \\ \boldsymbol{\alpha}\beta_0 &\equiv \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}'\boldsymbol{\mu}, \end{aligned} \tag{5}$$

with $\beta_{\perp}'\beta = \mathbf{0}$ and $\alpha_{\perp}'\alpha = \mathbf{0}$. Such a representation will be *tested for*, and cointegration (C_{fp}) used as a minimal requirement to be satisfied for the revision process of $I(1)$ variables to be meaningful.

In this framework, we can also consider *efficiency* (EF_{fp}) as a necessary and sufficient condition for p_t to yield an efficient forecast of f_t in the MSPE sense

$$EF_{fp} \Leftrightarrow E\left(f_t|p_t, F_{t-1}, P_{t-1}\right) = E\left(f_t|p_t\right), \tag{6}$$

where $F_{t-1} = \{f_{t-j},\ j = 1, 2, \ldots\}$ and $P_{t-1} = \{p_{t-j},\ j = 1, 2, \ldots\}$. In such a case, current preliminary data also contain all information available in past values of final and preliminary data.

Defining now the ratio of conditional covariance between Δf_t and Δp_t to the conditional variance of Δp_t as $\omega_{fp} = \sigma_{12}/\sigma_{22}$, we can exploit the properties of the conditional expectations for a bivariate normal random variable (see e.g., Spanos, 1986, Ch.15) to manipulate the VAR representation (1) to yield the model for f_t conditional on p_t:

$$\left(a_{11}(L) \Leftrightarrow \omega_{fp} a_{21}(L)\right) f_t = \left(\omega_{fp} a_{22}(L) \Leftrightarrow a_{12}(L)\right) p_t + \left(\mu_1 \Leftrightarrow \omega_{fp}\mu_2\right) + u_t. \tag{7}$$

EF_{fp} holds if and only if no lags of f_t or p_t are relevant in the conditional model. Recalling the definition of $a_{ij}(L)$, we can say that EF_{fp} is equivalent to

$$f_t = \omega_{fp} p_t + \left(\mu_1 \Leftrightarrow \omega_{fp}\mu_2\right) + u_t. \tag{8}$$

Hence, EF_{fp} corresponds to cointegration, C_{fp}, and to having uncorrelated error correction terms $z_t = f_t \Leftrightarrow \mu_1 \Leftrightarrow \omega_{fp}(p_t \Leftrightarrow \mu_2)$, $t = 1, 2, \ldots, T$, which are two proeprties easily tested for.

Note that using on first differences the same algebra as before we get

$$\begin{aligned} \left(b_{11}(L) \Leftrightarrow \omega_{fp} b_{21}(L)\right) \Delta f_t &= \left(\omega_{fp} b_{22}(L) \Leftrightarrow b_{12}(L)\right) \Delta p_t \\ &+ \left(\omega_{fp}\alpha_2 \Leftrightarrow \alpha_1\right) z_{t-1} + \left(d_1 \Leftrightarrow \omega_{fp} d_2\right) + u_t, \end{aligned}$$

or

$$\Delta f_t = c + \omega_{fp}\Delta p_t + \psi z_{t-1} + u_t.$$

As a consequence, in the presence of cointegration, the inclusion of the error correction term is essential and hence studying data revision properties based on

relationships such as $\Delta f_t = \gamma_0 + \gamma_1 \Delta p_t + \epsilon_t$ is prone to an omitted variable bias. This is a point often overlooked in the literature.

Last, a necessary and sufficient condition permitting preliminary data to be *unbiased* (U_{fp}) forecasts of the corresponding final data is

$$U_{fp} \Leftrightarrow E\left(f_t | p_t, F_{t-1}, P_{t-1}\right) = p_t. \tag{9}$$

Rewriting the equilibrium relationship (4) as $f_t = \Leftrightarrow\beta_0 \Leftrightarrow \beta_1 p_t + z_t$, we have *zero-mean revision errors* (ZMRE_{fp}) when $(\beta_0,\ \beta_1) = (0,\ \Leftrightarrow 1)$. Thus, $U_{fp} \Leftrightarrow EF_{fp} \cup ZMRE_{fp}$.

As remarked before, we are allowing for the presence of a constant in the VECM, $\mathbf{d}$, and in the cointegration relationship, β_0, and this requires special attention in the testing procedure. Although a joint test is possible, we will report the outcome of an alternative two-step test for ZMRE, as we deem it more informative: we first test whether $\beta_1 = \Leftrightarrow 1$, that is, whether revision errors are stationary; then we test for $\mu_1 = \mu_2$ given that, conditional on $\beta_1 = \Leftrightarrow 1$, $\beta_0 = 0$ if and only if $\mu_1 = \mu_2$, hence testing whether the revision errors have a zero mean.

All properties and testing procedures are summarized in Table 1.

2 Preliminary and Final Data on US M1

As an illustration of how the properties just described can be assessed we will refer to the relationship between preliminary and final data on M1 for the US.[1] We study the period from January 1973 to August 1995, using monthly seasonally adjusted data (the only data available to us). We are aware of the possible limitations deriving from the use of these data since the adjustments in the seasonal coefficients which accompany the overall revisions might have an impact on the outcome of the tests (Kavajecz and Collins, 1995). Yet, the type of distortions found by these authors in seasonally adjusted data do not appear in our results, once cointegration and the search for a correct dynamic specification are properly inserted into the analysis.

Table 1

Summary of Properties and Testing Procedures

Property	VECM	Hypothesis	Test Used
Cointegration	Unrestr. (2)	$\text{rank}(\mathbf{A}(1) = 1)$	Johansen
Efficiency	Restr. (3)	z_t uncorrelated	LM for uncorrelation
Zero-Mean Rev. Errors	Restr. (3)	$(\beta_0, \beta_1) = (0, -1)$	LR for $\beta_1 = -1$ and Wald for $(\mu_1 = \mu_2 \mid (\beta_1 = -1))$ alternatively, joint LR test
Unbiasedness		EF + ZMRE	

For the sake of brevity, we will provide evidence just on the bivariate relationship between preliminary and final data (i.e., the first published and the latest available data). The results for the other relationships (first revision-preliminary and final-first revision, cf. Figures 1c and 1d) will be summarized below and are available upon request. [2]

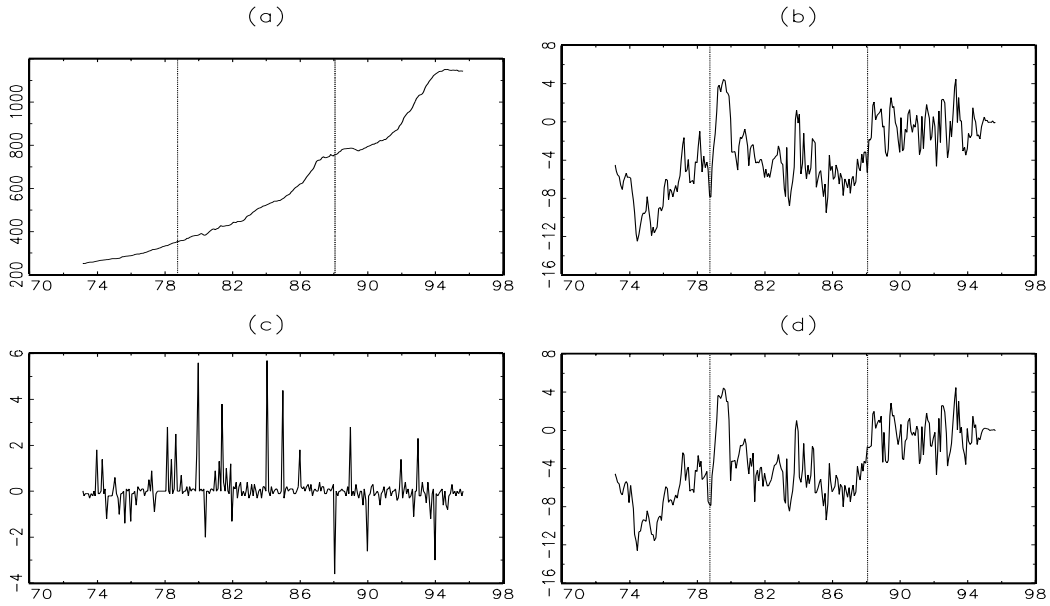

Figure 1: US M1 - Provisional and Final Data: 1973:01 - 1995:08
(a) Final Data; (b) Final-Preliminary; (c) Revised-Preliminary
(d) Final-Revised. Vertical bars correspond to the breaks.

The behavior of the levels of provisional and final data is such that one would not distinguish one from the other from a graphical point of view, hence we report only the latter in Fig. 1a; however detailing the difference between final

and preliminary data, $f_t \Leftrightarrow p_t$, one can see that it oscillates around a value quite different from zero (Fig. 1b) and that it seems to behave differently across sub-periods. The strategy we follow is therefore to formally test for the presence of a break in $f_t \Leftrightarrow p_t$ at an unknown time period, by applying the tests by Andrews (1993) and Andrews and Ploberger (1994) for a null hypothesis of joint constancy of all parameters (cf. the Appendix for a brief description of the tests) in a simple model (an AR(1) with a constant) to detect the possible breaks and then apply the detailed analysis and testing to each sub-period in a VAR context, subjecting the estimates to further stability testing.

The results show that the null of no break on the entire period is strongly rejected (p-values for the joint constancy test statistic are 0.008 (SupLM test), 0.002 (ExpLM), 0.0004 (AveLM)); the estimated break point is December 1987. However, visual inspection of the first sub-period suggests that a further break might have occurred. This is indeed the case, judging from the results of the tests on the null hypothesis of no break computed on the period January 1973 – December 1987 (p-values: 0.017 (SupLM), 0.004 (ExpLM), 0.001 (AveLM)), with an estimated break point at July 1979.[3]

The interpretation of these break points can only be tentative: the only major monetary events around these dates are the change in the operating procedures by the Fed between October 1979 and September 1982 documented by various authors (e.g., Hamilton, 1988) whereby interest rate targeting was abandoned in favor of money supply, Greenspan was appointed Chairman of the Fed in August 1987 and the Stock Exchange crashed in October 1987. It is generally recognized that starting from that period the Fed has put in place an increasingly transparent announcement procedure, and possibly paid more attention to the quality of preliminary data. Note that the timeline provided by Kavajecz (1994) suggests that no definitional changes in money supply occurred at or around the break points isolated here. Whether the latter are due to an technical improvement in the data production process or to a deliberate policy change is therefore still an open question.

We will then conduct our analysis on the three sub-samples separately, namely, January 1973 to July 1979, August 1979 to December 1987, and January 1988 to August 1995. We start by presenting our results for the cointegration tests in

Table 2:[4] the hypothesis of the existence of one cointegrating vector is accepted in all periods. Hence the basic requirement for the data revision process is satisfied.

Table 2

Cointegration Tests: Preliminary and Final Data

Sample	H_0: rank=p	λ-max	95% C.V.	Trace	95% C.V.
73:01-79:07	$p = 0$	*24.3*	19.0	*34.3*	30.1
	$p \leq 1$	9.9	12.3	9.9	12.3
79:08-87:12	$p = 0$	*30.2*	14.1	*30.55*	15.4
	$p \leq 1$	0.35	3.8	0.35	3.8
88:01-95:08	$p = 0$	*26.4*	14.1	*26.8*	15.4
	$p \leq 1$	0.33	3.8	0.33	3.8

Critical values from Osterwald-Lenum (1992).
Trend included (restricted to lie in the cointegrating space) in the first sub-sample.

We can now test whether the restrictions on the cointegrating relationships apply across the subperiods, testing first the hypothesis of zero-mean revision errors. Looking at the first row in Table 3 (period 1973:01-1979:07), we see that, although the hypothesis $\beta_1 = \Leftrightarrow 1$ is marginally accepted, this is so in the presence of a trend in the cointegrating space and hence there is no interest in testing the second requirement for level unbiasedness, i.e., whether $\beta_0 = 0$. This is not surprising in view of the upward trend in the revision errors for the period at hand (cf. Figure 1b).

Table 3

Relationship between Preliminary and Final Data Tests on the Data Revision Properties

Sample	$(\beta_1) = (-1)$	$(\mu_1 = \mu_2 \mid (\beta_1) = (-1)$	Level Efficiency
1973:01-1979:07	*3.65 [0.06]*	—	*66.48 [0.00]*
1979:08-1987:12	0.98 [0.32]	*15.68 [0.00]*	*80.42 [0.00]*
1988:12-1995:08	0.04 [0.85]	0.36 [0.55]	*27.96 [0.00]*

p-values in square brackets;
Test for $\beta_1 = -1$) is a LR $\sim \chi^2(1)$;
Test for $\mu_1 = \mu_2 | \beta_1 = -1$ is a Wald $\sim \chi^2(1)$;
Test for efficiency is an LM test for uncorrelation of the error correction term $\sim \chi^2(6)$.

As for the second sub-sample, the tests point to the stationarity of the revision

errors, although the hypothesis of no constant term in the error correction term is rejected (in agreement with the inspection of Figure 1b). The third and last sub-sample is characterized by zero mean stationary revision errors.

Efficiency is rejected for all periods (last column of Table 3) suggesting that the preliminary data do not summarize all the informational value contained in previous preliminary and final data and that lagged values should be taken into account as well.[5]

Finally, for the other bivariate relationships, cointegration is present in all cases; for $f_t \Leftrightarrow r_t$ (graphed in Figure 1d) we obtain the same results as for $f_t \Leftrightarrow p_t$ (no ZMRE for the first two periods, no unbiasedness, no efficiency), whereas for $r_t \Leftrightarrow p_t$ (graphed in Figure 1d) all desirable properties are satisfied.[6]

3 Ex ante analysis

When considering expectations formation, the actual content of the currently available information set becomes a binding constraint. In view of the results obtained in the *ex post* analysis, both past final and provisional data appear to be relevant for the determination of final data and hence should all be included in the information set used for forecasting.

Let us indicate with $f_{t|t+1}$ the optimal (in a MSPE sense) forecast of final data for period t made at $t+1$ (after data for p_t and r_{t-1} have been published). We have

$$\begin{aligned} f_{t|t+1} \equiv E(f_t|I_{t+1}) = E(\Delta f_t + \Delta f_{t-1} + \ldots + f_{t-k}|I_{t+1}) = \\ E(\Delta f_t|I_{t+1}) + E(\Delta f_{t-1}|I_{t+1}) + \ldots + f_{t-k}, \end{aligned} \tag{10}$$

where we have assumed

$$I_{t+1} = \{p_j, r_{j-1}, \ldots, f_{j-k}, \quad j = k+1, \ldots, t\}, \tag{11}$$

that is, at time $t+1$ we lack values of $f_t, \ldots, f_{t-k+1}$. Our focus will then be on the elements of (10), $E(\Delta f_{t-j}|I_{t+1})$, $j = 0, \ldots, k \Leftrightarrow 1$, noting that they are equivalent to $E(\Delta f_t|I_{t+1+j})$, $j = 0, \ldots, k \Leftrightarrow 1$.

The starting point will then be the joint analysis of (f_t, r_t, p_t) and the derivation of a conditional model for Δf_t under the different assumptions about the

information set. Let us assume, as before, that there exists a VAR(q) generating process $\mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{e}_t$ where, this time, $\mathbf{y}_t = (f_t, r_t, p_t)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$, and $\mathbf{e}_t \sim i.i.d.N(\mathbf{0}, \boldsymbol{\Sigma})$. To simplify matters, we will assume that $k = 2$ and rewrite this model as a suitable error correction model:

$$\Delta\mathbf{y}_t = \dot{-}\boldsymbol{\alpha}\mathbf{z}_{t-2} \Leftrightarrow \mathbf{G}\Delta\mathbf{y}_{t-1} + \mathbf{H}(L)\Delta\mathbf{y}_{t-3} + \mathbf{e}_t \tag{12}$$

where $\mathbf{G} = \{g_{ij}\} = (\mathbf{I} \Leftrightarrow \mathbf{A}_1)$ and

$\mathbf{H}(L) = (\mathbf{H}_0 + \mathbf{H}_1 L + \ldots + \mathbf{H}_{q-3}L^{q-3})$, $\mathbf{H}_i = \Leftrightarrow\sum_{j=i+3}^{q}\mathbf{A}_j$, which differs from the usual EC representation because the error correction terms, $\mathbf{z}$, appear with lag two.

From (12) we need to derive a conditional EC model for Δf_t to be used for forecasting purposes. We have:

$$\begin{aligned} \Delta f_t &= a_1\Delta f_{t-1} + u_t \\ &\quad +\omega_{13}\Delta p_t + c + \gamma_1 z_{1t-2} + \gamma_2 z_{2t-2} \\ &\quad +a_2\Delta r_{t-1} + a_3\Delta p_{t-1} + \mathbf{h}(L)'\Delta\mathbf{y}_{t-3} \\ &\equiv a_1\Delta f_{t-1} + u_t + K_{t-1} \end{aligned} \tag{13}$$

where $\omega_{13} = \sigma_{13}/\sigma_{33}$, and, in an obvious notation, $\gamma_1 = \alpha_{11} \Leftrightarrow \omega_{13}\alpha_{31}$, $\gamma_2 = \alpha_{12} \Leftrightarrow \omega_{13}\alpha_{32}$, $c = d_1 \Leftrightarrow \omega_{13}d_3$, $a_1 = g_{11} \Leftrightarrow \omega_{13}g_{31}$, $a_2 = g_{12} \Leftrightarrow \omega_{13}g_{32}$, $a_3 = g_{13} \Leftrightarrow \omega_{13}g_{33}$, $u_t = e_{1t} \Leftrightarrow \omega_{13}e_{t3}$, while $\mathbf{h}(L)$ is a 3×1 vector whose elements are $\mathbf{h}_i(L) = h_{1i}(L) \Leftrightarrow \omega_{13}h_{3i}(L)$, $i = 1, 2, 3$.

Care is to be exerted in this case, since such a model contains Δf_{t-1}, itself unknown[7] at time $t+1$. Therefore, we need to substitute this unknown value with its expression in terms of known variables and the lagged error term. The outcome is a model which is notationally cumbersome and involves an MA(1) error term, as is usual with more than one-step-ahead forecasts. Thus, by backward substitution of Δf_{t-1} in (13), we find the model which will be used to forecast in practice:

$$\Delta f_t = a_1^2\Delta f_{t-2} + e_t + K_{t-1} + a_1 e_{t-1} + a_1 K_{t-2}. \tag{14}$$

Notice that if $a_1 = 0$ (a condition to be verified in practice), the model reduces to $\Delta f_t = K_{t-1} + e_t$, which implies an uncorrelated error term and a simpler forecasting structure.

One period later, at $t+2$, we still do not know the value of Δf_t, but we have additional information in the form of Δp_{t+1}, Δr_t, z_{1t-1}, and z_{2t-1}. From an empirical perspective, then, in order to compute $\Delta f_{t|t+2}$, we will add these variables to the list of regressors in the forecasting conditional model (14). Recall that we do not need to substitute for Δf_{t-1}, since its value is known at $t+2$.

We will perform here an *ex ante* forecasting exercise separately for each of the three subperiods (1973:01-1977:07, 1979:08-1985:12, 1988:01-1993:07). We construct different conditional error correction models (CM) for Δf_t, leaving a horizon of 24 periods for each sub-sample to perform a forecast comparison evaluation (*à la* Diebold and Mariano, 1995, cf. the Appendix) between the outcome of our conditional models (re-estimated each time) and a number of alternatives described below.

Starting from the trivariate restricted VECM, we have derived the implied CM for Δf_t and Δf_{t-1}, by deleting irrelevant regressors. The resulting models retained have a very different specification across sub-samples.[8] For $\Delta f_{t|t+1}$ we have

$$\begin{aligned}
\mathbf{73:01} \Leftrightarrow \mathbf{77:07} \quad &: \quad \text{Constant}, \Delta p_t, r_{t-1} \Leftrightarrow p_{t-1} \\
\mathbf{79:08} \Leftrightarrow \mathbf{85:12} \quad &: \quad \text{Constant}, \Delta p_t, r_{t-1} \Leftrightarrow p_{t-1}, \Delta r_{t-1}, \\
&\qquad f_{t-2} \Leftrightarrow r_{t-2}, \Delta f_{t-i}, i = 4,5 \\
\mathbf{88:01} \Leftrightarrow \mathbf{93:07} \quad &: \quad \text{Constant}, \Delta p_t, \Delta r_{t-1}, f_{t-2} \Leftrightarrow r_{t-2}
\end{aligned}$$

while for $\Delta f_{t|t+2}$ we have

$$\begin{aligned}
\mathbf{73:01} \Leftrightarrow \mathbf{77:07} \quad &: \quad \text{Constant}, \Delta r_t, r_t \Leftrightarrow p_t \\
\mathbf{79:08} \Leftrightarrow \mathbf{85:12} \quad &: \quad \text{Constant}, r_t \Leftrightarrow p_t, \Delta r_t, f_{t-1} \Leftrightarrow r_{t-1}, \Delta p_{t-3} \\
\mathbf{88:01} \Leftrightarrow \mathbf{93:07} \quad &: \quad \text{Constant}, \Delta p_{t+1-i}, i = 0,1, \Delta r_{t-i}, i = 0,\ldots,4 \\
&\qquad \Delta f_{t-i}, i = 1,\ldots,4, f_{t-1} \Leftrightarrow r_{t-1}
\end{aligned}$$

As we can see, the list of retained regressors in the model for Δf_t is a subset of K_{t-1} in expression (13), from which we can infer that $a_1 = 0$ and hence that we do not need to consider MA(1) disturbances.[9]

The CMs are estimated recursively to generate one-step-ahead forecasts of the variations of money supply for each sample point in the forecasting period, based on information available at time $t+1$ and $t+2$, $\Delta f_{t|t+1}$ and $\Delta f_{t|t+2}$, respectively.

In order to provide a meaningful evaluation for our model, we will contrast its performance against simple alternative forecasts of final values, constructed

from available data at each successive time period. First, we construct *naïve* forecasts (N) based on the most recent, although provisional, data for Δf_t, hence $p_t \Leftrightarrow r_{t-1}$ at time $t+1$ and $r_t \Leftrightarrow f_{t-1}$ at time $t+2$. If p_t and r_t were unbiased and efficient for f_t, N would be the minimum MSPE forecasts. Then, in view of our results about the bias of r_t for f_t, we suggest to correct $r_t \Leftrightarrow f_{t-1}$ by the mean revision error (evaluated recursively to mimic real time updating), labeling the corresponding forecasts as *naïve corrected* (NC). Finally, we label as *purist* (P) the forecasts based on final data alone, i.e., discarding available provisional information. Assuming a characterization of the process for M1 as a random walk with drift, the purist forecasts will be the means of the first differences of f_t, i.e., $(\overline{f \Leftrightarrow f_{-1}})|(t+1)$, computed up to time $t \Leftrightarrow 2$ for I_{t+1} and $(\overline{f \Leftrightarrow f_{-1}})|(t+2)$, computed up to time $t \Leftrightarrow 1$ for I_{t+2}. Thus, we have three competing forecasts, CM_1, N_1, P_1 for $t+1$ and four, CM_2, N_2, NC_2, and P_2, for $t+2$.

We summarize the results in Table 5 where we report both absolute and quadratic loss criteria (Mean Absolute Prediction Error – MAPE – and Mean Squared Prediction Error – MSPE), and the Diebold and Mariano (D-M) statistic S_1 (cf. the Appendix), a test for the significance of the difference in MAPEs and MSPEs between our conditional model CM_i, (i=1,2) as a benchmark against each of the alternative models. The sign of the estimated S_1 indicates whether the corresponding forecast is better (positive sign) or worse (negative sign) than the benchmark in each subset. The p-values reported should thus be judged against a one sided alternative (significantly better or significantly worse).

The results show a marked difference of performance in terms of MAPE and MSPE between the forecasts at time $t+1$ and those at time $t+2$. As one would expect, the values are generally lower in the second set, as it is based on a larger information set (although not significantly so – the D–M test results are not reported). No major differences arise from the use of an absolute or a quadratic criterion. For the first set of forecasts (at time $t+1$), the conditional model provides predictions which are as good as the naïve predictions (once significance is taken into account) but significantly better than the purist ones (particularly so for the second and third subperiods). For the second set (at time $t+2$), the conditional model has a better performance than the naïve forecast (although the latter improves once the correction for the systematic discrepancy between

revised and final figures is inserted), and even more so for the forecasts based on final figures alone.

Table 5

Forecast of $\Delta \mathbf{f}_t|t+1$ and $\Delta \mathbf{f}_t|t+2$

Forecasting Diagnostics

Period 1977:08 - 1979:07 – h=24							
	$\mathbf{I_{t+1}}$			$\mathbf{I_{t+2}}$			
Diagnostic	CM_1	N_1	P_1	CM_2	N_2	NC_2	P_2
MAPE	1.07	0.85	1.24	1.03	4.81	1.20	1.24
D-M S_1		1.36	-0.95		-6.83	-0.60	-1.20
p-values		(0.17)	(0.34)		(0.00)	(0.55)	(0.23)
MSPE	1.64	1.43	2.52	1.50	26.53	2.53	2.51
D-M S_1		0.53	-1.49		-5.19	-1.31	-1.71
p-values		(0.60)	(0.14)		(0.00)	(0.19)	(0.09)
Period 1986:01 - 1987:12 – h=24							
MAPE	1.17	1.08	4.82	1.07	5.63	0.97	4.79
D-M S_1		0.46	-5.93		-10.26	0.39	-6.83
p-values		(0.64)	(0.00)		(0.00)	(0.70)	(0.00)
MSPE	2.22	1.45	32.46	1.73	33.26	1.46	32.07
D-M S_1		0.98	-3.63		-8.20	0.35	-3.80
p-values		(0.33)	(0.00)		(0.00)	(0.72)	(0.00)
Period 1993:09 - 1995:08 – h=24							
MAPE	0.75	0.90	4.02	0.65	0.67	0.72	3.98
D-M S_1		-0.52	-5.82		-0.10	-0.30	-5.82
p-values		(0.60)	(0.00)		(0.92)	(0.77)	(0.00)
MSPE	0.97	1.73	22.26	0.58	1.00	1.11	21.85
D-M S_1		-1.16	-4.12		-1.01	-1.27	-4.09
p-values		(0.25)	(0.00)		(0.31)	(0.20)	(0.00)

Models for	$\Delta \mathbf{f}_t$ at $t+1$	$\Delta \mathbf{f}_t$ at $t+2$
CM Conditional Model:	$\Delta f_{t\|t+1}$	$\Delta f_{t\|t+2}$
N Naïve:	$p_t - r_{t-1}$	$r_t - f_{t-1}$
NC Naïve Corrected:		**N** - $(\overline{r-f})|(t+2)$
P Purist:	$(\overline{f-f_{-1}})|(t+1)$	$(\overline{f-f_{-1}})|(t+2)$

Overall, therefore, the results signal that the use of extra information suggested by the ex post analysis leads to some improvement in the prediction, and that

purist forecasts based on final data have a significantly worse overall performance.

Finally, we should comment on the fact that our results are obtained under the assumption that final data are available with a two-period delay. Lifting this hypothesis would presumably have two major effects on our comparisons: the first is that the proper conditional models change since the relevant final values in the information set would need to be substituted with intermediate revisions. In this respect, the stylized facts about the lesser degree of importance in successive revisions suggest that the empirical evidence should not vary by much. The second effect is that the purist forecast would further worsen its performance as it relies on a smaller information set.

4 Conclusions

The unavailability of timely error-free data can have serious consequences in empirical work and in the process of expectations formation. In this paper we have suggested an econometric framework to analyze the relationship between provisional and final data taking into account the nonstationarity of the processes, and to test for some desirable properties of the data production process.

The empirical application of this procedure was performed on US money supply data (M1). Our results show that the period from 1973:01 to 1995:08 was characterized by two endogenously detected structural breaks. The in-sample study of the characteristics of the data indicates that cointegration between provisional and final data is always present, and that this has strong consequences for the specification of the most suitable model describing such a relationship. As one would expect, the relationship between preliminary and revised data is the strongest and exhibits most of the desirable properties. With the notable exception of the first sub-sample, the cointegration analysis shows that the difference between provisional and final data is stationary, but only in the last period around a mean of zero.

To complement the ex post analysis, we have also suggested a suitable procedure for deriving a conditional model for first differences of final data which only includes real time information. Such a model was used to forecast unavailable money supply final data on the basis of currently available information. The

results show that the conditional model provides some improvements (although they are not always significant) in forecasting money supply movements over alternatives which rely on natural transformations of provisional data or of past final data.

The conditioning set on which we operate is admittedly the smallest possible. Other improvements and richer models could be investigated by including other variables of interest. This is left for future research.

References

[1] Anderson, R.G. and K.A. Kavajecz, 1994, A Historical Perspective on the Federal Reserve's Monetary Aggregates: Definition, Construction and Targeting, *Federal Reserve Bank of St. Louis Review*, 76, 1-31.

[2] Andrews, D.W.K., 1993, Tests for Parameter Instability and Structural Change with Unknown Change Point, *Econometrica*, 61, 821-56.

[3] Andrews, D.W.K., and W. Ploberger, 1994, Optimal Tests when a Nuisance Parameter is Present Only under the Alternative, *Econometrica*, 62, 1383-1414.

[4] Barro, R.J., and Z.Hercowitz, 1980, Money Stock Revisions and Unanticipated Money Growth, *Journal of Monetary Economics*, 6, 257-267.

[5] Boschen, J.F., and H.I.Grossman, 1982, Tests of Equilibrium Macroeconomics Using Contemporaneous Monetary Data, *Journal of Monetary Economics*, 10, 309-333.

[6] Diebold, F.X., and R.S.Mariano, 1995, Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-63.

[7] Diebold, F.X. and G.D. Rudebusch, 1991, Forecasting Output with the Composite Leading Index: A Real-Time Analysis, *Journal of the American Statistical Association*, 86, 603-10.

[8] Doornik, J.A. and Hendry, D.F., 1994, *PCFIML 8.0: An Interactive Econometric Modelling of Dynamic Systems*, London: International Thomson.

[9] Engle, R.F., and C.W.J.Granger, 1987, Cointegration and Error Correction: Representation, Estimation and Testing, *Econometrica*, 55, 251-276.

[10] Hakkio C. S. and M. Rush, 1989, Market Efficiency and Cointegration: An Application to the Sterling and Deutschemark Exchange Markets, *Journal of International Money and Finance*, 8, 75-88.

[11] Hamilton, J.D., 1988, Rational-expectations Econometric Analysis of Changes in Regime: An Investigation of the Term Structure of Interest Rates, *Journal of Economic Dynamics and Control*, 12, 385-423.

[12] Hamilton, J.D., and G. Perez-Quiros, 1995, What do the Leading Indicators Lead?, UCSD DP 95/22.

[13] Hansen, B.E., 1997, Approximate Asymptotic P Values for Structural–Change Tests, *Journal of Business and Economic Statistics*, 15, 60-7.

[14] Johansen, S., 1995, *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*, Oxford: Oxford University Press.

[15] Kavajecz, K.A., 1994, The evolution of the Federal Reserve's Monetary Aggregates: A Timeline, *Federal Reserve Bank of St. Louis Review*, 76, 32-66.

[16] Kavajecz K.A. and S. Collins, 1995, Rationality of Preliminary Money Stock Estimates, *Review of Economics and Statistics*, 77, 32-41.

[17] Maravall A., and D. Pierce, 1986, The Transmission of Data Noise into Policy Noise in U.S. Monetary Control, *Econometrica*, 54, 961-979.

[18] Osterwald–Lenum, M., 1992, A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistic, *Oxford Bulletin of Economics and Statistics*, 54, 461-71.

[19] Patterson, K.S., and Heravi, 1991, Data Revisions and the Expenditure Components of GDP, *The Economic Journal*, 101, 887-901.

[20] Spanos, A., 1986, *Statistical Foundations of Econometric Modelling*, Cambridge, Cambridge University Press.

Appendix

Tests for Structural Stability

The Andrews (1993) and Andrews and Ploberger (1994) tests are conveniently summarized by Hansen (1997), whose GAUSS code we used to compute the test statistics and the approximate p-values. The tests consider the break point k as unknown which makes the testing procedure for no breaks subject to the so-called problem of having a nuisance parameter identified only under the alternative hypothesis and makes the asymptotic distribution of a nonstandard type. The three test statistics, SupLM, ExpLM and AveLM are derived from the repeated computation of a Lagrange Multiplier test F_T for given k, making then k vary within a certain range (in our case its position relative to T was varied between 0.15 and 0.85, as is customary). We have:

$$\begin{aligned} SupLM &= \sup_{k_1 \leq k \leq k_2} F_T(k) \\ ExpLM &= \log\left(\frac{1}{k_2 \Leftrightarrow k_1 + 1} \sum_{k=k_1}^{k_2} \exp\left(\frac{1}{2} F_T(k)\right)\right) \\ AveLM &= \frac{1}{k_2 \Leftrightarrow k_1 + 1} \sum_{k=k_1}^{k_2} F_T(k). \end{aligned}$$

Test for Forecasts Performance Comparison

The forecast comparison test proposed by Diebold and Mariano (1995) is motivated by the need to provide a general framework for predictive accuracy even in the presence of non–normality or autocorrelation in the forecast errors $\hat{e}_{it}$ and $\hat{e}_{jt}$ from two competing models. These errors are transformed through a $g(\cdot)$ function and a difference d_t is constructed as $d_t = g(\hat{e}_{it}) \Leftrightarrow g(\hat{e}_{jt})$. Since the distribution of $\bar{d} = \sum_{t=1}^{T} d_t/T$ is given by $\sqrt{T}(\bar{d} \Leftrightarrow \mu) \rightarrow \mathcal{N}(0, 2\pi f_d(0))$, where $2\pi f_d(0)$ is the variance expressed in the frequency domain which can be estimated as the value of the spectral density at frequency zero. The (asymptotically normally distributed) test statistic for the null hypothesis of no difference between the two competing forecasts (i.e., $\mu = 0$) is

$$S_1 = \frac{\bar{d}}{\sqrt{\frac{\widehat{2\pi f_d(0)}}{T}}}.$$

Endnotes

[1] For an exhaustive description of the data production for this aggregate, see Anderson and Kavajecz (1994) .

[2] The VAR analysis was implemented in PCFIML 8.0 (Doornik and Hendry, 1994); the remaining computations were performed in GAUSS 3.2.

[3] We are fully aware of the fact that the tests were not designed for sequential analysis of break points; however, the estimated p-values are so low that we feel fairly confident that the null of no break in the subperiod can still be rejected. Note also that no further break points could be detected.

[4] The full details on parameter estimation and on residual diagnostics are omitted and are available upon request. The lag length selected for each subperiod were, respectively, 4, 4, and 3, following a general-to-specific modelling strategy based on their significance from a Wald test and the non-correlation of the residuals. No major problems were detected with autocorrelation, heteroskedasticity, normality, and ARCH. In particular, the Andrews (1993) and Andrews and Ploberger (1994) tests on the residuals accept the null hypothesis of joint constancy of the parameters on each subperiod.

[5] Full details are available upon request, as well as for the multivariate version of the testing procedure confirming the bivariate analysis.

[6] Also Δr_t is unknown at time $t+1$ and we cannot condition on its value.

[7] The existence of a cointegrating relationship which involves preliminary and revised data only allows us to consider $p_{t-1} \Leftrightarrow r_{t-1}$ as a regressor.

[8] This is also confirmed by the residual diagnostics which show no problems.